Fall 11-13-2019

Spontaneous Recombination of Short RNAs to Increase Length and Complexity in Prebiotically Plausible Conditions

Benedict Arthur Smail
Portland State University

Spontaneous Recombination of Short RNAs to Increase Length and Complexity in Prebiotically Plausible Conditions

by

Benedict Arthur Smail

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Chemistry

Dissertation Committee:
Dirk Iwata-Reuyl, Chair
Albert Benight
John J. Perona
Christof Teuscher

Portland State University
2019

Abstract

The RNA World hypothesis is an influential theory of how life arose on earth which posits that the earliest life forms carried out essential activities with RNA catalysts instead of proteins. Substantial evidence to support this theory comes from among others, the fact that the ribosome is a ribozyme, the existence of small RNA catalysts capable of high activity, and self-assembling ribozymes such as the *Azoarcus* ribozyme. Additional support has been provided by the directed evolution of various ribozymes and the reputed prebiotic synthesis of nucleotides and small RNA oligomers. However, it has long been challenging to show how long RNA catalysts could have emerged from prebiotic syntheses or be formed by nonenzymatic processes. Here, I suggest that energetically neutral RNA-RNA recombinations could have created longer oligomers from short precursors generated from nonenzymatic prebiotic syntheses. I show that small RNAs less than 20 nucleotides long are capable of recombination by up to three unique mechanisms and confirm that the activity is intrinsic to the RNA. I further show that the reaction profile changes under different conditions, including pH and magnesium concentration, and that some products may have distinctly different structures. Finally, I show that the reaction improves with time and can be reproduced with similar RNAs according to sequence or structure. These results demonstrate that RNA-RNA recombination is an inherently plausible means of diversification, elongation, and enhancement for small, abiotically generated oligomers, and indicate that a capacity for spontaneous RNA-RNA recombination is a fundamental property of all RNA molecules.

Dedication

To Mom, Dad, Irene, Gregory, and all my family near and far who supported me and took this path before me.

To the unbridled forces of nature whence we came, and to which we all return.

Acknowledgements

Dirk Iwata-Reuyl

Committee members:

Albert Benight

John Perona

Christof Teuscher

Coauthors:

Bryce Clifton

Ryo Mizuuchi

Niles Lehman

I would like to thank current and former lab members of the Lehman lab and Iwata-Reuyl lab, Portland State University for supporting me throughout my time here, and all others who gave support, encouragement, or who put their faith in me.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The origin of life is one of the great unsolved mysteries of science. Darwin's theory of evolution, and the subsequent discovery of genetics, and later, phylogenetics[1,2], convincingly show that life on earth has evolved over time by natural selection. The process of natural selection is both spontaneous and inevitable; in any population of organisms facing any selection pressure, mutations in their genomes from generation to generation will inevitably render some proportion more fit than others and those genes will become prevalent in the population.

The principle of common descent is one of the cornerstones of evolutionary theory. Living organisms are not designed, rather they evolved previously from other organisms. Various subfields of evolutionary theory, such as comparative embryology, physiology, and anatomy reveal the fundamental similarities of related organisms. The discovery and application of phylogenetic analysis has made it possible to properly characterize and identify related species. Even more remarkably, phylogenetic analysis can be used to show that life likely evolved from a common ancestor over 3.8 billion years ago. Yet, natural selection has not existed for all eternity; the earth itself is not much older than 4.5 billion years, when a cloud of space dust coalesced into the wet, watery planet we know today as Earth[2].

The mechanism of natural selection is genetic change, which is enforced by high-fidelity protein DNA polymerases that copy the genome during reproduction of every living thing[3,4]. The protein polymerase is one of the most fundamental units of life – it preserves the blueprint of successful organisms and passes it down to offspring. Nonetheless, like all other components of living things, these high-fidelity polymerases,

as well as the genetic polymers they copy, have themselves evolved over time and their origin is one of the great conundrums of evolutionary theory. Evolution explains how organisms change, at a more granular level it explains how biomolecules change, but it struggles to explain how the reproduction machinery of an organism could have spontaneously appeared out of non-living chemicals.

The polymerase itself is a protein, a polymer of amino acids that when assembled and folded, takes a nucleic acid template and synthesizes a complement, the first step in reproduction of a genome. However, proteins do not merely arise out of nothing; a concentrated solution of amino acids might undergo some condensations to form short polypeptides, but the odds of a high-fidelity polymerase, hundreds of amino acids long, emerging from some chemically treated pool of amino acids are infinitesimally small. One cannot simply mix chemicals and create life. Like all proteins, the polymerase must be manufactured, and in nature, it is made in living organisms by the ribosome, an industrial-sized (among biomolecules) RNA-protein complex whose ribosomal RNA is the true agent of amino acid polymerization off of an RNA template[5].

1.1 RNA catalysis at the origins

The knowledge that RNA is the agent of amino acid polymerization[5] begs the question of where the RNA came from. In nature today, RNA is polymerized on a DNA template by RNA polymerase, another protein forged by the ribosomal RNA. Given the appropriate conditions and resources, the RNA polymerase and ribosome would appear to form a self-replicating set. One makes the other, which makes the other in turn, and natural selection can begin. Indeed, some of the simplest RNA viruses are just this: a

piece of RNA whose transcription by the ribosome releases an RNA polymerase that transcribes anew the RNA that led to its own creation[6].

Which then came first, the ribosome or the RNA polymerase? The principle of natural selection would tell us that they must have evolved from simpler precursors. Did they evolve together, from nothing? Did one emerge before the other? Were they possibly seeded on Earth by random processes? Were they perhaps dropped on Earth by aliens, or even gods? The theory of panspermia[7] seems unlikely; space is too hostile, especially the unfiltered cosmic radiation emitted by the sun. For aliens to seed life, they presumably would have needed to arise themselves, leaving the same questions unresolved. And there is no evidence for gods.

It seems more likely than not that life must have arisen on earth. By following natural selection, one can reasonably conclude that peptidyl transferase and polymerase activity have themselves evolved. However, mere reductionism cannot reveal the origins of these molecules. For one thing, transcription of both proteins and RNA is intimately dependent on various adapters and cofactors, as well as "food" molecules. Take away tRNA and the ribosome cannot translate; without nucleotide triphosphates, the polymerase cannot transcribe. The problem becomes even more difficult when the food runs out – what makes the nucleotide triphosphates or the tRNA in the first place?

Another problem with both reductionism and natural selection to explain the origins of peptidyl transfer or polymerase activity is that there is simply a limit on how small these catalysts can be. The necessary functions of the enzymes, gathering the substrates and the required cofactors, catalyzing the reaction, and releasing the finished molecule, require a molecule large enough to form a structure encompassing all these

abilities, and this molecule will invariably have a sequence space so high that only an infinitesimal fraction of it can be explored by the best physical processes.

Nevertheless, in spite of the difficulties with the size of the catalyst, one of the best, and indeed principal approaches to the origin of life has been to search for smaller and smaller catalysts[8]. The only question left then is which catalyst? Should one look for an RNA catalyst or a protein catalyst? The fact that RNA functions as both a genetic molecule and a catalyst makes it far more relevant for studying the origin of life than a protein or polypeptide catalyst[9], which cannot function as a genetic polymer, and thus cannot provide a mechanism of inheritance. In addition, the building blocks of RNA – nucleobases and nucleotides – can be formed under putative prebiotic conditions[10] but it is more difficult to form amino acids. Furthermore, since the discovery of ribozymes, it has been found that RNA catalysts can be quite small. The catalytic portion of the HDV ribozyme, responsible for hepatitis delta virus infection in humans, is a mere 78 nucleotides long[11], and there is a wide variety of small self-cleaving ribozymes 50 nucleotides or less in length[12].

1.2 RNA-catalyzed RNA polymerization

The capacity of RNA to act as a catalyst and genetic polymer implies that in theory, it should be possible to make a self-replicating RNA molecule; a molecule that can catalyze its own production. Indeed, ever since Gilbert's proposal of the RNA World in 1985[13], a considerable amount of effort has been expended on the search for an RNA polymerase ribozyme, which is widely considered both proof-of-concept of the RNA World and a possible means of early reproduction[14].

The foundation for these approaches first began with the *Tetrahymena* ribozyme shortly after its discovery in the 1980s[15]. By transcribing the *Tetrahymena* ribozyme without its 5′ exon, and adding a 3’ terminal guanosine to serve as the nucleophile, it was shown that the ribozyme could catalyze the sequential recombination of pentacytidine into polycytidine oligomers up to 30 nt in length[16]. The reaction is a disproportionation reaction in which one cytidine is attached to the end of pentacytidine to make hexacytidine, with tetracytidine as a leaving group (Figure 1). The reaction can turn over with the poly-cytidine products to make longer products.

In the case of poly-cytidine, the reaction is accomplished with the internal guide sequence of the *Tetrahymena* ribozyme, GGAGGG, which serves as the recognition site for poly-cytidine oligomers[15, 16]. However, modification of this site to contain different nucleotides (and thus different substrates) does not produce robust results[17]. It is also not possible to effectively separate the template from the ribozyme[18], meaning that the Tetrahymena ribozyme is a poor model for a primordial RNA replicase.

The search for an RNA polymerase ribozyme was begun in earnest by the discovery of the class I ligase in 1993[19]. Developed by Bartel and Szostak, the class I ligase was isolated from a pool of 220-nt length oligomers after 10 rounds of *in vitro* selection based on its ability to ligate a triphosphorylated substrate to its 3′ terminus (Figure 2)[19]. The success of the class I ligase prompted over 25 years of work to improve it to be a true RNA polymerase ribozyme (RPR). One recent iteration by Horning and Joyce (Figure 3)[20] can polymerize a tRNA in 0.07% yield, whereas another can polymerize RNAs of similar lengths to the RPR[21]. Nonetheless, the proposal of the RNA

polymerase ribozyme still faces some significant challenges to explain its possible role in either RNA-only lifeforms or its emergence from non-enzymatic, abiotic processes.

The most obvious problem with the RPR is that despite decades of research, there has not been much progress finding an RPR that can truly polymerize a copy of itself. Many factors contribute to this problem – first, the ribozyme must polymerize RNA nucleotides off of an RNA template which has to bind and then dissociate from both the product RNA and the enzyme. Second, the nucleotides used must be activated, ostensibly with prebiotically plausible activating agents. Third, the ribozyme must traverse significant regions of secondary structure – the class I ligase features a double-nested pseudoknot in its active site[22].

1.3 Abiotic generation of prebiotic RNA catalysts

In terms of prebiotic chemistry, a far more significant problem with the Bartel-Szostak RPR is its size; at roughly 200 nucleotides, the sequence space exploration required to find it from random abiotic processes is prohibitively high. Even with nucleotide activation, the spontaneous generation of ribozymes with the size and function of an RNA polymerase ribozyme has remained a challenge. The experiments of Ferris and Orgel showed that nucleotides preactivated with imidazole can be polymerized by montmorillonite clays into oligomers up to 55 nucleotides in length[23], with the bulk of the products falling within the range of 20-30 nucleotides. Even the longest oligomers in this very successful and heralded experiment are only about a third of the size of a typical class I ligase.

One approach to the size problem has been to truncate the RPR down to a minimal effective length, but this results in a ribozyme that is still quite large –

approximately 150 nucleotides – due to the size and critical structure of its active site. Another approach – repeating the original Bartel-Szostak experiment with shorter pools such as 100-nt or 150-nt pools – has never been reported in the literature. Still, these difficulties do not rule out an RPR as an engine of prebiotic life; it may be that 25 years is not long enough to evolve the best possible RPR, or that the proper conditions to assist ribozyme-catalyzed RNA polymerization haven't been discovered. It is also possible that the function of a hypothetical, primitive RPR was not to polymerize itself but to polymerize nucleotides or extend short templates into ribozyme-sized oligomers.

Many different chemical conditions, including under plausible prebiotic conditions, have been shown to produce short RNA oligomers from a variety of prebiotically plausible precursors. For example, Krishnamurthy used diamidophosphate to achieve production of poly-uridine oligomers[24]. By using wet-dry cycles in hot acidic conditions, Deamer was able to achieve the polymerization of poly-adenosine up to 100 nucleotides[25] and has also explored the efficacy of lipid-assisted RNA polymerization[26]. More recently, Holliger[27] used the hairpin ribozyme to extend oligomers with nucleoside 2'-3' cyclic phosphates in ice, and Szostak's group demonstrated primer extension with activated monomers and trimers[28]. However, all of these experiments fail to reach both the combined size and nucleotide diversity likely required for advanced catalysts.

It may be that relevant prebiotic conditions for long abiotic RNA polymerization or elongation simply haven't been found yet. On the other hand, the reality of nucleotide polymerization chemistry is that there is a high energy barrier to reacting, and nucleotides must usually be activated for any polymerization to occur. These activating agents may or may not be prebiotically plausible, but if they can be shown to be relevant, an activation

step would be necessary. Yet thus far, even activation with high-energy leaving groups such as imidazole[28] is not sufficient for formation of long polymers.

1.4 Relevant catalytic abilities of RNA

The longstanding difficulty of forming long RNA catalysts (>100 nt) by strictly abiotic means raises the question of whether there are ways to obtain those catalysts by mechanisms other than non-enzymatic polymerization of activated nucleotides. For example, it is well known that there are many small RNA catalysts, both in nature and selected from *in vitro* directed evolution. The highly active HDV ribozyme is roughly 78 nucleotides in length[11]; a minimized version of the hairpin ribozyme is 50 nucleotides long with a 14-nt substrate[29], and small hammerhead ribozymes are also approximately 50 nucleotides long[30]. *In vitro* selection has isolated even smaller ribozymes such as a 20-nt ligase ribozyme that attaches a 14-nt substrate[31], or the Yarus pentamer 5′-GUGGC[32], which, upon binding to its substrate 5′-GCCU can accelerate the self-aminoacylation of multiple amino acids activated with AMP.

It is also well known that short RNA oligomers can form transient tertiary structures, or trans complexes with low dissociation constants. This was established by Doudna and Cech, who broke apart the active site of the Tetrahymena ribozyme and were able to observe reformation of inactive, non-covalent tertiary structures[33]. The ability of RNA oligomers to form such trans complexes was exploited by Hayden and Lehman, who broke apart the Azoarcus ribozyme, a modified Azoarcus group I intron, and observed its formation into a catalytically active trans complex that catalyzed recombination of four Azoarcus ribozyme fragments as small as 39 nucleotides into the full-length covalent ribozyme[34]. This length, as well as the lengths of many ribozymes

from directed evolution experiments, is well within the range of lengths generated by various prebiotically plausible synthetic routes.

1.5 Transesterification and energetically neutral recombination

Of the many small ribozymes that have been discovered, either in nature or by *in vitro* selection, there is one property that nearly all have in common. This is the ability to catalyze a transesterification reaction that is typically initiated by the attack of an internal 2′-hydroxyl onto its adjacent phosphodiester to release a 5′ oligomer and form a 2′-3′ cyclic phosphate, or by attack of a 3′-hydroxyl on a scissile phosphate followed by release of the 5′ group to form a recombinant product. The former represents the mechanism of the HDV[11], hairpin[29], and hammerhead[30] ribozymes among others, while the latter is the usual mechanism of group I introns[35]; this chemistry has also been the inspiration for directed evolution of the class I ligase and other ligase ribozymes to attack a triphosphate moiety instead of a phosphor-ester[19].

The ability of even very small ribozymes to accelerate transesterification reactions portends the ability of short RNAs to form recombinant products. This RNA recombination, as it is known, has additional virtues that make it prebiotically plausible. Instead of a high-energy leaving group, the attack of a ribose hydroxyl on a phosphate requires a nucleoside, nucleotide, or polynucleotide leaving group, and is thermodynamically favorable compared to nonenzymatic polymerization of unactivated substrates[36]. The reaction can take place in water (metal ions are likely required) and its principle limitation is its reversibility. However, in a prebiotic RNA world without optimized sequences, a greater degree of reversibility may have been useful for sequence space exploration and swapping of information.

The scientific literature does contain a few examples of putative prebiotic RNA recombination. Di Mauro's group reported both the recombination and ligation of poly-C and poly-G oligomers non-enzymatically polymerized from cyclic GMP and CMP[37]. Their proposed mechanism is a consequence of uneven base-pairing, a terminal, overhanging 3′-hydroxyl attacks either a 5′ phosphate or the last phosphodiester bond at the 5′ end to form a ligation product or recombination product respectively (Figure 4). However, the substrates lack inherent nucleotide diversity and their proposed models feature reactions without metal ions and with water as a leaving group in a ligation reaction, a reaction that seems both kinetically and thermodynamically implausible.

Perhaps one of the more intriguing models of recombination from small linear RNAs of relevant prebiotic lengths is the reaction originally developed by Vlassov[38] (Figure 5). In this reaction, two identical, short RNA strands 16 nucleotides long are brought in close proximity by a template strand, or splint, and recombine to form a 28mer product. The splint, oriented from 3' to 5' binds to both strands with a unique region of full Watson-Crick complementarity for each. The poly-A tail of the top strand does not bind to the splint and instead drifts free above the double-stranded complex. In the first step of the recombination reaction, the poly-A tail is spontaneously (and specifically) cleaved, leaving a terminal guanosine with a 2′,3′-cyclic phosphate. The guanosine remains hydrogen-bonded to a cytosine in the splint and in the second step, the cyclic phosphate is attacked by the nearby 5′-hydroxyl of the second strand above the splint, forming either a 2′-5′ or 3′-5′ linkage between the two top strands. Digestion of the product strand, a 28-mer, with ribonuclease T1, which cleaves only the 3′-5′ bond at the 3′-end of a guanosine, has

demonstrated that a majority of the linkages formed in the cleavage and ligation reaction are of the 2´-5´ variety[38].

Although the splint in this model is too small to form a significant tertiary structure and does not catalyze a chemical reaction in the traditional sense of an enzyme, it has all the properties of a multiple-turnover ribozyme. It substantially accelerates the cleavage and ligation reaction that occurs at negligible levels in its absence; it may dissociate from the complex upon completion from the reaction, it is substrate-specific according to its hydrogen bonding, and it is neither consumed nor altered in the reaction. Moreover, it is possible for the reaction to happen a second time, producing a 40-nt strand from the product 28-mer. It is possible that structure affects the reaction in a limited way: the 11-nt base-paired region is enough to form an A-form alpha helix which may promote cleavage of the overhanging tail.

The basic features of the Vlassov reaction imply that nearly every short RNA strand could potentially be a ribozyme if given the right substrates – a strand binding part of the 3′ end of the splint and a strand binding part of the 5′ end that displaces part of the other strand. I will refer hereafter to this mechanism as “alpha” recombination; being the first example of small-RNA recombination whose reliability is not in doubt, which is characterized by the setup in Figure 5: a splint strand oriented from 3′-5′, a substrate strand which binds approximately half (or part) of the splint at the 3′ end, a substrate strand binding approximately half (or part) of the splint at the 5′ end, and the displacement of a small portion of the first substrate, which has no Watson-Crick complementarity to the splint, and that is extruded and specifically cleaved prior to ligation of the two substrates.

1.6 New mechanisms of length expansion via RNA recombination

Although there is evidence that RNA or RNA building blocks can be formed under relevant prebiotic conditions there is a substantial gap in our understanding of how small RNA oligomers could have become large RNA catalysts. Therefore, the primary goal of this research has been to investigate whether spontaneous RNA recombination is a viable means of length expansion among small, prebiotically plausible RNAs. Although the discovery of recombination over a splint by Lutay et al.[38] and the proposed reaction of Di Mauro[37] suggest that recombination is possible even for very small oligomers, it has never been investigated in a systematic fashion. After the discovery that RNAs stored in the freezer for long time periods expand their lengths[39], a study of RNA recombination with colleagues Bryce Clifton and Ryo Mizuuchi demonstrated that random pools of 16mers will spontaneously recombine to generate longer sequences that in some cases exceed twice the size of the original starting material. However, although high-throughput sequencing of such recombinant products confirms their size, it has been unclear what specific recombination mechanisms are operative in random RNA pools.

Building on past research, I sought to design specific, short RNA oligomers 16 nucleotides in length or less that will undergo recombination during self-incubation or joint incubation. The use of fixed sequences ensures that any recombinant products discovered during sequencing can be analyzed for specific mechanisms. Here, with the specific design of short synthetic oligomers 16 and 13 nucleotides in length, **R16** and **H13** respectively, I demonstrate up to three new mechanisms of recombination by short RNAs to produce longer and more complex recombinant products. Furthermore, in contrast to the primarily pyrimidine substrates of Vlassov[38] and single-base

polynucleotides of Di Mauro[37], the results presented here will show that recombination is not limited by nucleotide composition, geometry, or consecutive base-pairing. Finally, I support these findings with chemically explicit computer simulations of alpha-recombination, which show that random cleavage and ligation events over time can result in the gradual accumulation of longer oligomers, structures, and ultimately ribozymes that can feed back and further accelerate recombination into larger oligomers. Taken together, my work shows that RNA recombination is a plausible means of achieving longer and more complex RNAs from short precursors generated by non-enzymatic abiotic processes.

1.7 Dissertation contributions

In this dissertation, I investigate whether RNA recombination can be a means of length expansion and complexification among small RNAs strands. My work is important because there is a gap in our scientific understanding of how large RNA catalysts, which would have been relevant for protobiotic life, could have been formed from small precursors at the origin of life. It is known in the scientific literature that there are plausible chemical pathways to short RNA oligomers in prebiotic chemistry, but it has not been possible to generate long RNA oligomers with the size and diversity necessary for advanced catalysts.

I investigated RNA recombination using both *in vitro* studies on synthetic RNA oligomers and sequence-explicit computer simulations of recombination. My *in vitro* studies examined the recombination of designed RNA oligomers into longer recombinant products by multiple novel mechanisms. My computer simulations provide the first simulation of RNA recombination containing full sequence information for every

oligomer in the simulation, and introduce the concept of replenishment as a means of initiating the origin of life by recombination.

The detailed contributions of my dissertation are as follows.

- I designed two small RNA oligomers, **R16** and **H13**, which were 16 and 13 nucleotides long respectively, to undergo recombination when incubated by themselves or with each other.
- I showed that the self-incubations of **R16** and **H13** produce multiple recombination products and provided strong evidence for up to three distinct mechanisms of recombination. One is a cleavage and ligation mechanism termed the *alpha-prime* process, in which cleavage and ligation occur over a 3-nt bulge. This type of RNA recombination is a new mechanism that has never been previously reported in the literature. A second mechanism is a one-step transesterification initiated by an internal 2′ hydroxyl on a phosphodiester bond to form a branch, and a third is a one-step attack of a terminal 3′ hydroxyl on a phosphodiester bond to form a linear molecule.
- I found that a joint incubation of **R16** and **H13**, which was designed to test a mechanism proposed by Di Mauro[37], did not produce products, indicating that this proposed mechanism is not facile.
- I demonstrated that the 5′-OH of **R16** and **H13** is the nucleophile in the alpha-prime process, and can be effectively blocked by 5′ phosphorylation.
- I provided sequencing evidence to support the existence of the *beta* recombination product, which results from the attack of a terminal 3′-OH onto another molecule of **R16** to form a linear RNA product.

- I provided strong evidence of branched RNAs among the recombination products of **R16** by ligating adapters to the 3′ ends, resulting in products that could only have been formed from ligation of the adapters to multiple 3′ terminals.
- I generalized the alpha-prime process of **R16** and **H13** to the sequence space of all 16mers, and found that a subset of 16mer sequence space (0.17%) is capable of undergoing self-recombination by a similar process. Based on my generalization, I designed and procured the oligomers F16 and B16, and demonstrated that both oligomers undergo self-recombination to form larger products.
- I wrote a nucleotide-explicit computer simulation in which every oligomer had a fixed sequence, polarity, preactivation with cyclic phosphates or not, simple secondary structure, and structural classification.
- In my simulation, I designed an algorithm to select RNA strands at random from a pool and form alpha-prime structures using strict criteria for base-pairing, leaving groups, and consecutive base pairs. Oligomers that formed alpha-prime structures were subject to site-specific cleavage and ligation with certain probabilities to form recombinant products. I tested the algorithm's accuracy at forming alpha-prime structures by simulating recombination using the oligomers of Lutay et al.[38] and confirmed that the algorithm finds plausible alpha-prime structures with 100% accuracy.
- I used my algorithm to simulate recombination on a flat distribution of RNA oligomers 16 nucleotides long, and an exponential distribution of oligomers ranging from 5-17 nucleotides long, and showed that RNA recombination leads to

a redistribution of the starting material that includes a substantial proportion of longer strands.

- I showed that if transient ribozymes are able to feed back on the rates of cleavage and ligation in my simulation, the overall rate of recombination is accelerated and produces a larger quantity of longer strands.
- I demonstrated in my simulation that replenishment of a pool of RNA oligomers undergoing recombination allows the long-term accumulation of structure in the pool, providing a conceptual framework for how recombination can lead to the formation of complex RNA catalysts.

Chapter 1 Figures:

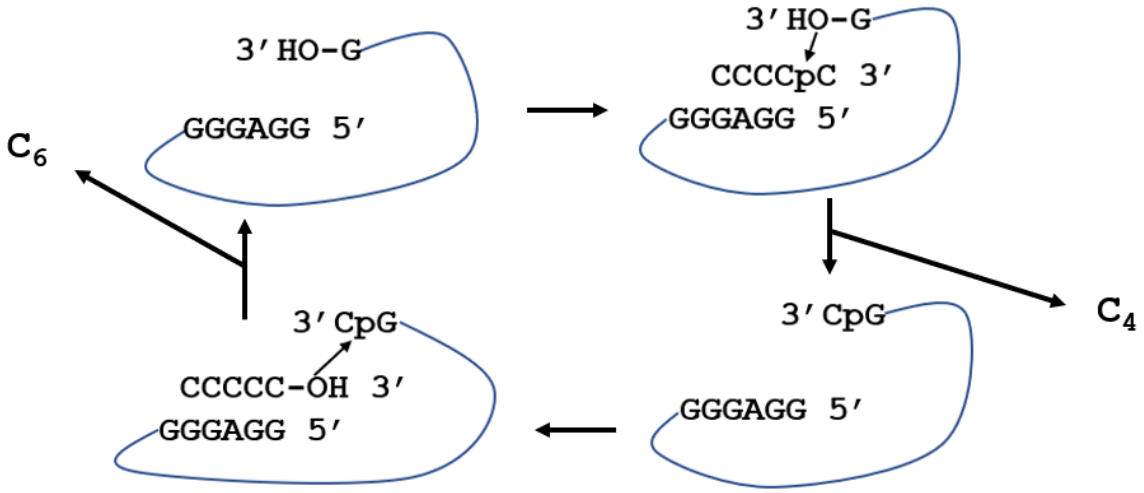

Figure 1: The disproportionation reaction of the Tetrahymena L-19 IVS RNA. The ribozyme incorporates pC at the end of pentacytidine to form C_6 and releases C_4. The reaction can turn over to form progressively longer products. Illustration based on (16).

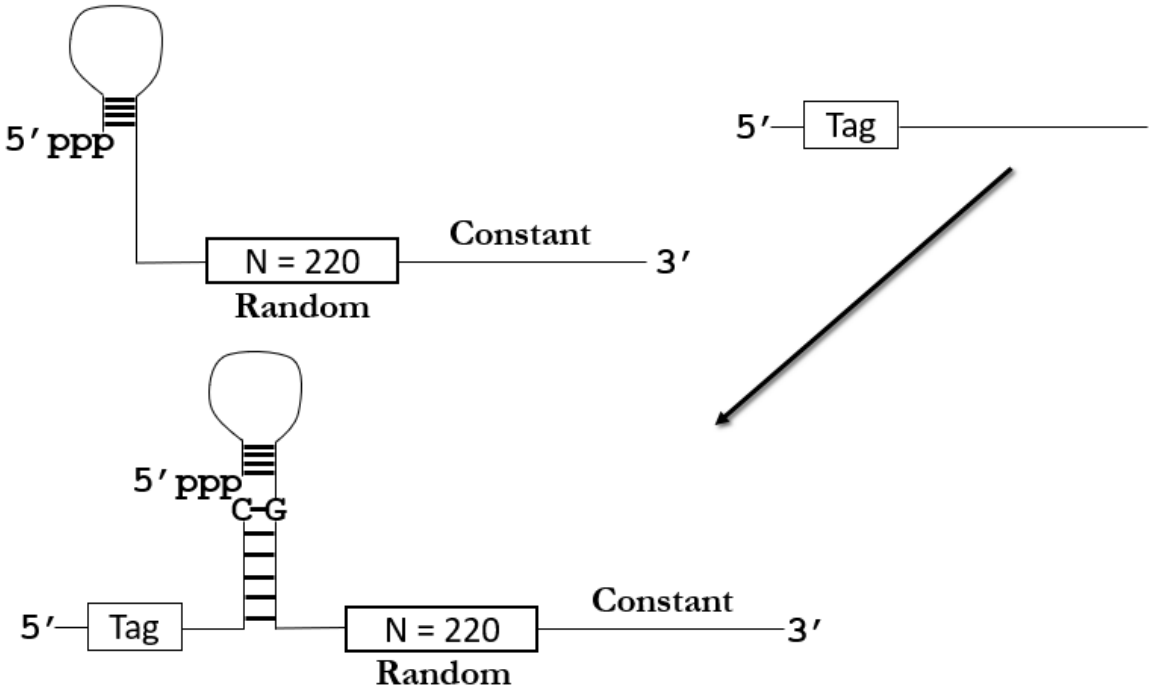

Figure 2: Scheme for directed evolution of an RNA polymerase ribozyme by Bartel and Szostak. Illustration based on (19).

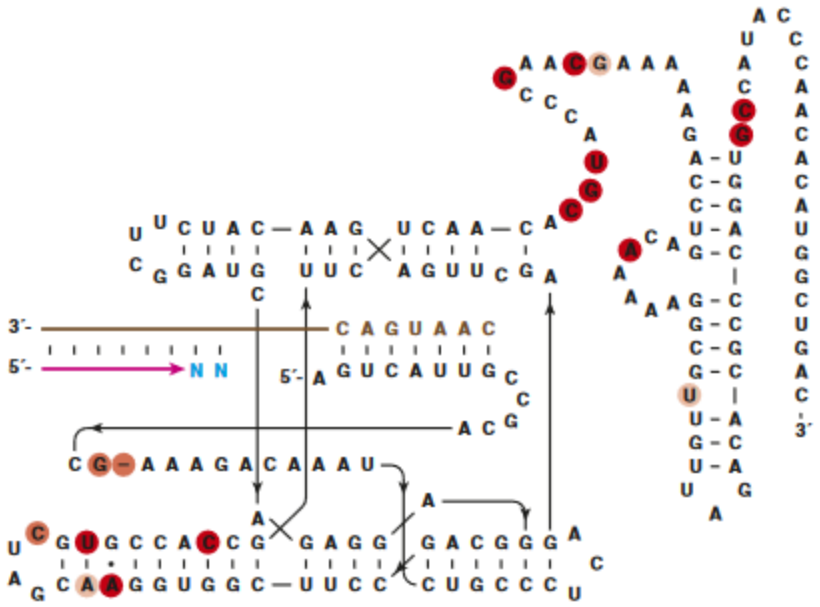

Figure 3: The 24-3 RNA polymerase ribozyme of Horning and Joyce that can polymerize a tRNA in 0.07% yield. Figure from (20), reprinted with permission.

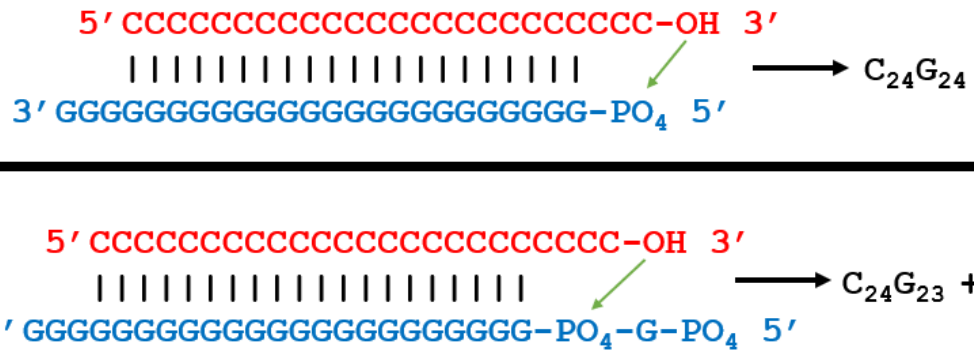

Figure 4 (top): Ligation model proposed by Di Mauro. A terminal 3′-OH of C_{24} is proposed to attack the 5′ phosphate of a complementary strand, G_{24}, that is unevenly bound. An alternative version (bottom) proposes attack of the same terminal hydroxyl on the penultimate phosphor-ester in a recombination reaction, also in water and without metal ions. Illustration based on (37).

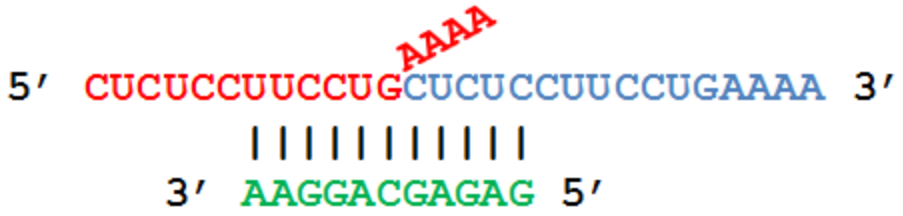

Figure 5: The cleavage and ligation scheme from Lutay et al.[38], showing two 16-nt oligomers splinted over a third 11-nt RNA strand. The extruded AAAA tail drifts free above the double-stranded complex and is specifically cleaved to yield a 2′-3′ cyclic phosphate on the guanosine residue. The 5′-OH of the 5′ cytosine on the adjacent strand may attack the cyclic phosphate and ligate to form a new 28mer. The net reaction is 16 + 16 = 28 + 4. Figure based on description in (38).

Chapter 2: Experimental Methods

R16 (5′-CGUACCGUUGCAUUUG) and **H13** (5′-CUGCAACGGUACG) were purchased as HPLC-purified and dried RNA pellets from either TriLink Biotechnologies (TLB) or Integrated DNA Technologies (IDT). The pellets were rehydrated in 100 µM EDTA to chelate any metal ions that had persisted in the solid phase synthesis and the concentration of oligomers was determined by absorbance measurements. RNAs were gel purified to eliminate impurities, rehydrated in 0.1 mM EDTA at a concentration of ~100 µM for the stock solution, and stored at -80°C until use.

For reactions, RNA stock solutions were thawed to room temperature, and an appropriate amount was taken for each reaction before being returned to the -80°C freezer. RNAs were diluted in water to approximate working concentrations, heated to 65°C for two minutes, and cooled to room temperature with the addition of buffer and 1.0 M magnesium chloride. For pH values of 8.0 and below, the buffer Tris was used; for pH values 8.7 and above, CHES was used. The final reaction contained 50 mM buffer and 100 mM magnesium chloride, while the concentration of RNA was typically 40-80 µM. These conditions will be referred to hereafter as "standard conditions."

RNAs were incubated for various time periods in a PCR machine or at room temperature. Incubation times ranged from one hour to 30 days, with the bulk of incubations done for 5-7 days. For standard conditions, reactions were cycled every three hours between 22°C and 0°C for the duration of the reaction, and isothermal trials were performed at room temperature (ca. 22°C). Upon completion of incubation, reactions were ethanol precipitated to remove excess salt, rehydrated in a gel-loading mix containing 96% formamide, 10 mM EDTA, and bromophenol blue, and stored at -80°C

until gel loading. RNAs were electrophoresed through either an 8%, 15%, or 22% denaturing polyacrylamide gel with 8M urea at 600, 800, and 1000-1200 volts respectively until the bromophenol blue dye line was 1-2 inches from the bottom of the gel.

RNA products were visualized with radioactive labeling by P_{32} or staining with either SYBR™ Gold or SYBR™ Green. For radioactive labeling, RNAs were labeled either before reacting or upon completion. For labeling prior to incubation, RNA was radiolabeled with Optikinase (USB, discontinued) or T4 Polynucleotide Kinase (NEB), and subject to an organic extraction and ethanol precipitation prior to rehydration in water. A small quantity of labeled RNA (~2-4 pmol) was then added to reactions containing the working concentration of RNA. For radiolabeling after completed reactions, reacted RNA was ethanol precipitated and rehydrated in water and kinase buffer before being radiolabeled and electrophoresed on a 15% polyacrylamide gel.

To stain RNA with either SYBR™ Gold or SYBR™ Green (Thermofisher) concentrated stain was diluted according to the manufacturer's instructions. Gels electrophoresed with product RNA were soaked in the dark with diluted stain for 20-40 minutes (SYBR™ Gold) or 40-60 minutes (SYBR™ Green). Visualization of labeled or stained RNAs was carried out on a Typhoon Trio+ imager.

2.1 Sanger and high-throughput sequencing

RNA recombination products were analyzed either with Sanger sequencing or high-throughput sequencing (HTS) on the MiSeq at the core facilities of Oregon State University. In order to prepare RNAs for sequencing, recombinant RNA products were excised from 15% polyacrylamide gels and subject to an overnight crush-and-soak with

sodium acetate and EDTA to separate them from the gel. After ethanol precipitation, RNAs were ligated to a 3′ DNA adapter with T4 RNA ligase 2, truncated KQ227 (NEB). The 3′ adapter contained a dideoxy tail to eliminate byproducts or double ligations and had previously been adenylated according to a procedure adapted for Illumina sequencing[40]. RNAs were subsequently phosphorylated at the 5′ end with either T4 polynucleotide kinase (NEB) or Optikinase (USB, discontinued), and a 5′ adapter was attached with T4 RNA ligase 1. The resulting products were reverse transcribed with Superscript IV reverse transcriptase and amplified over 25-30 PCR cycles with Taq DNA polymerase.

PCR products from the above preparation were gel-purified on a 1% or 2% agarose gel, excised, and isolated using the Qiagen DNA purification kit. For HTS sequencing, products were diluted to approximately 2 nM and shipped on ice to Oregon State's core facilities for qtPCR and sequencing. For Sanger sequencing, DNA products were concentrated to 7-8 µL and prepared with the CloneJet PCR Cloning kit (Thermofisher). Ligation products were mixed with competent cells and grown overnight on luria broth agar with 0.1 (units) ampicillin. Colony PCR products were treated with the Thermofisher Big Dye terminator kit, ethanol precipitated, and sent to Oregon Health and Science University for Sanger sequencing.

2.2 Transcription with T7 RNA polymerase

For transcription of modified **R16** with a G at the first nucleotide instead of C, we procured the DNA template 5′-CAAATGCAACGGTACCTATAGTGAGTCGTATTA with 2′-methoxy substitutions on the first two nucleotides of the 5′ end. DNA was annealed to the T7 promoter at a concentration of 20 µM each and 100 mM NaCl by

heating it to 90°C for two minutes in a thermocycler and allowing it to cool to room temperature. Five µL of annealed DNA was mixed with 10 µL 10X transcription buffer (containing spermidine), 7.5 mM of each nucleotide triphosphate (NTP), and 1 µL of inorganic pyrophosphatase (2 units/µL). Magnesium was added to produce a final concentration of 36 mM, or six millimolar greater than the total NTP concentration, and the reaction was topped with water and 2 µL T7 RNA polymerase to a final volume of 100 µL. Reactions were performed for 18-24 hours at 37°C, followed by the addition of 2 µL DNase I and a further one hour of incubation. Transcribed RNA was purified with the addition of water and 1/10 volume sodium acetate prior to a double organic extraction with acid phenol and single organic extraction with chloroform-isoamyl alcohol. After organic extraction, the RNA was then precipitated with five or ten volumes of 100% ethanol and gel purified on a 15% polyacrylamide gel.

For transcription of a hammerhead ribozyme that self-cleaves to **R16**, the DNA template 5′-CAAATGCAACGGTACGGACGGTACCGGGTACCGTTTCGTCCT CACGGACTCATCAGCCGGTCTCCCTATAGTGGAGTCGTATTA was used, which had methoxy modifications on the last two nucleotides of the 5′end. Transcription of this template was designed to produce a hammerhead ribozyme that released **R16** (the last 16 nucleotides) with an invariant cleavage reaction after the CAG sequence just prior to **R16**. The procedures for transcription of this template were identical to those described above, except a double gel purification was used to isolate putative **R16**. However, as discussed later, cleavage of this hammerhead produces heterogenous pieces and overall cleavage is poor. Future work could be performed to optimize the hammerhead and its cleavage for best recovery of true **R16**.

Chapter 3: Spontaneous recombination of short, synthetic oligomers

In order to examine whether RNA recombination is a viable means of diversification among short RNA strands, two RNA strands were designed to test as many variations of proposed RNA recombination mechanisms as possible. **R16** (5′-CGUACCGUUGCAUUUG) was chosen to match the size of the oligomers used by Lutay et al.[38] with the inclusion of 50% G-C content, whereas **H13** (5′-CUGCAACGGUACG) was chosen to be complementary to **R16** save for a single nucleotide at the 5′ end of H13. The 5′ termini of **R16** and **H13** were specifically designed with a cytosine after experiments with random RNAs revealed an enrichment of cytosine at the 5′ end[38]. In addition, the 3′ end of each is terminated with guanosine, since this is the critical nucleotide for group I intron activity[35]. The predominant duplex (inferred by maximal hydrogen bonding) formed with **R16** and **H13** contains a three-nucleotide overhang on the 3′ end of R16. Attack of the terminal 3′ hydroxyl of **R16** on the last phosphoester bond at the 5′ end of **H13** may produce a hairpin (Figure 6) similar to the mechanism proposed by Pino et al[37].

R16 and **H13** can each form several distinct secondary structures by base pairing either to themselves or each other. Both **R16** and **H13** can form a twisted self-duplex with two distinct regions of base pairing (Figure 7). An additional duplex can be formed by **R16** in which some base pairs are shifted downstream (Figure 8). In addition to the fully complementary duplex of **R16** and **H13**, both oligomers can produce a triplex in which two RNAs are splinted over a third with at least 8 consecutive base pairs (Figure 9). This setup, in which there is presumably a long 3′ tail over the base-paired region, could produce a ~22-nt oligomer via an alpha reaction. Another possible triplex can be

created with a one-nucleotide bulge in the splint strand (Figure 10). It is possible that other conformations with non-canonical Watson-Crick pairing may exist for both the self-duplexes and mixed duplexes, but an exhaustive molecular dynamics treatment fell outside the scope of the project.

Finally, **R16** and **H13** can each form a fully self-templating triplex (Figure 11). These triplex complexes loosely resemble the secondary structure of the deoxyribozyme-catalyzed branch reaction of Silverman (Figure 12)[41, 42], except with a 3-nucleotide bulge and smaller overhangs. The **R16** self-templating triplex features a 4-nt 3′ overhang and a 1-nt 5′ overhang spanning the 3-nucleotide bulge in comparison to the Silverman deoxyribozyme reaction, which had 6-nt on the 3′ end and 2-nt with a triphosphate moiety on the 5′ end spanning a 15-nt bulge. The triplex was designed primarily to inhibit attack of the 5′ hydroxyl on the adjacent strand, a key feature of the alpha reaction, but this will be shown not to be the case.

3.1 Results of recombinant experiments

An 8-day self-reaction of **R16** gives 7-8 clearly visible products (Figure 13); for **H13**, the 8-day incubation gives four clearly visible products (Figure 14). These products can be further classified into distinct sets using the proposed secondary structures of **R16** and **H13** and high-throughput sequencing (HTS) results of recombinant products (Table 1). For **R16**, there are three bands at 28, 29, and 30 nucleotides (Figure 13) whose HTS products are commensurate with the self-templating triplex (e.g. Figure 11). These products are designated as *alpha-prime* products to reflect their similarity to the alpha reaction, but with a 3-nt bulge spanning the reaction site.

In the **R16** self-reaction, there is a band at 31 nucleotides whose size and sequence (Table 1) is consistent with the self-templating triplex but the sequence cannot be explained by an alpha-like mechanism. We have designated this product as *beta*, and it may be the result of attack of the terminal 2′ or 3′-hydroxyl of one **R16** molecule onto the first phosphor-ester bond of another **R16** molecule positioned over the bulged splint.

There are two distinct regions of recombinant products which resolve on the gel at more than twice the size of the initial input molecules. The lesser of these regions contains multiple bands at approximately 34-36 nt and is fainter; the uppermost band migrates at roughly 38 nt and is more distinct. These bands are termed the *gamma* and *gamma-prime* bands respectively (Figure 13). No sequences of these sizes appeared in any of three distinct HTS runs that we performed, including a 75-base-paired end read, and we propose that these RNAs are branched RNAs.

In the case of the **H13** self-reaction, there are two distinct products at 25 and 27 nt (Figure 14). The 25 nt product is well-represented in HTS results (Table 2), which also reveal a similar product at 24 nucleotides in which one of the base-paired nucleotides has been cleaved off prior to ligation. These products are likely the result of the self-templating triplex in which an alpha-prime reaction similar to **R16** can occur, except with only two significant products due to the one-nucleotide guanosine tail. The 27-nt product does not manifest in the HTS results and as above, we propose that this is a gamma-like mechanism that produces a branched RNA.

For both **R16** and **H13**, to confirm that the products we observed on the gel are the products seen in our HTS data, we compared the relative intensity of the linear bands to the quantitative results of the sequencing (Figure 15). We found that a majority of

products are 29 nucleotides long, followed by a lesser amount at 30 nucleotides, while the amounts for the 28-nt product and beta are both smaller. These results parallel the results on our gels and indicate that sequencing products are consistent with those products. In addition, we attempted to measure the reaction yield of recombinant products by comparing the intensity of the recombinant bands to the intensity of the residual starting material. From these results, we find that for **R16**, the rough reaction yield for each linear band is 0.12%, 1.8%, and 0.59% for the alpha-prime 28, 29, and 30 bands respectively, and 0.23% for the beta band. For the gamma and gamma-prime regions, it is difficult to distinguish the individual bands so we measured the intensity for the entire region, resulting in 2.28% for the gamma region and 2.9% for the gamma-prime region. For **H13**, the reaction yield was 8.97% for the alpha-prime 25 and 6.78% for the gamma region. This is a considerable increase over the yields of **R16**, and while there is some variation in the intensity, it is not unreasonable to propose a 5% yield for the alpha-prime 25mer of **H13**.

Additional products for both **R16** and **H13** can be seen with longer incubations. A 10-day incubation of **R16** produces additional minor products at roughly 43 nt and 22-25 nt (Figure 16). The uppermost band could be the result of the alpha-prime reaction happening a second time, which is known to happen for the alpha reaction. The lower bands could be the result of base-induced self-cleavage or of recycling, in which molecules with cleaved tails react by the alpha-prime mechanism to form shorter alpha-prime products. For **R16**, the band at 22 nt could be the result of an alpha reaction (Figure 17) but this sequence was not detected in the HTS data. For **H13**, the 24-nt

product in the HTS data manifests well after longer timepoints and a second gamma product is visible as well (Figure 14).

The incubation of **R16** with **H13** under our standard conditions gives a different result than the self-incubations. When the concentrations of each oligomer are exactly equimolar (as determined by absorbance measurements) the self-reactions appear to be completely inhibited as no visible bands are apparent (Figure 18). In fact, the joint reaction does not produce any significant products, even ones that could be attributed to the type of reaction proposed by Di Mauro, in which the terminal, overhanging 3′ guanosine attacks the phosphor-ester bond of the unpaired 5′ nucleotide to form a hairpin (e.g. Figure 6).

We were unable to observe a reaction product consistent with a Di Mauro-type reaction under a wide range of experimental conditions, despite the fact that the duplex for the Di Mauro reaction should be the most predominant complex due to the stability imparted by its 12 base pairs. Still, it cannot be concluded with certainty that the Di Mauro reaction is impossible, since **R16** and **H13** do not contain the original sequences of the Di Mauro reaction. In the absence of visible product bands, we did not attempt any sequencing either. Nonetheless, it seems clear that even if the Di Mauro reaction is possible, it is clearly not facile.

3.2 R16 and H13 self-incubations under different reaction conditions

To further investigate the self-incubation reactions of **R16** and **H13**, they were carried out over a range of pH values and magnesium chloride concentrations. With magnesium chloride held constant at 100 mM, **R16** and **H13** reactions were compared side-by-side at pH values 7.0, 8.0, 9.0, and 10.0 (Figure 19). In both cases, there is an

increased reaction yield for the alpha-prime products from pH 7.0 to pH 9.0, along with increased degradation of the RNA at higher pH values.

We next examined the **R16** and **H13** reactions separately at the same pH range (7.0 to 10.0) but with steps every 0.5 pH units, and the results of the pH gel electrophoresis reveal the optimal pH values for each set of reactions. For **R16**, the middle gamma bands are strongest at pH 7.0, 7.5 and 8.0, while the gamma-prime band is strongest only at pH 8.0 (Figure 20). The **R16** alpha prime bands are not visible at pH 7.0 and are only faint at pH 7.5, consistent with a two-step mechanism involving base-catalyzed cleavage, which will be discussed in the next chapter. Both the alpha-prime and beta bands appear to be optimal at higher pH values; at pH 9.0 and 9.5, these bands are the most intense even though there is significant degradation of the RNA at these pH values. The pH dependence of the reactions implies that at least three distinct mechanisms are occurring: 1) the gamma reaction at lower pH values, 2) the gamma-prime reaction at pH 8.0, and 3) the alpha-prime and beta reaction at pH 9.0.

In the case of **H13**, pH variance from 7.0 to 10.0 reveals that the gamma band is most intense at pH 7.0-8.5 (Figure 21). Interestingly, whereas the alpha-prime bands of **R16** were not present at pH 7.0, the alpha-prime band of **H13** is visible at all tested pH values (Figure 21), although it increases significantly in intensity up to pH 8.5. Higher pH values appear to decrease the yield of both **H13** products. This is one of many observations that the alpha-prime reaction of **H13** is more robust than its counterpart in **R16**.

With pH held constant at 8.0 (the optimal pH to visualize all bands together), magnesium variance from 0-100 mM reveals that for **R16**, the alpha-prime and beta

bands require at least 50 mM magnesium chloride for significant product formation (Figure 22). The **R16** gamma bands are faintly visible at 0 mM magnesium, but are substantially enhanced in the presence of at least 5 mM magnesium. The omission of magnesium chloride and addition of 100 mM potassium chloride appears to permit the gamma reaction, but the alpha-prime and beta bands are not detected in the absence of magnesium.

Concomitant with the results of the pH experiments, the magnesium experiments reveal a more robust alpha-prime reaction for **H13** than for **R16**. Both the **H13** alpha-prime and gamma products are visible in as little as 5 mM magnesium while the intensity of the alpha-prime band increases with magnesium concentrations up to 100 mM.

In general, a majority of the **R16** and **H13** self-reaction products increase with time, with the gamma bands being the exception. To assess this at shorter times, the reaction was investigated over a total of 8 hours, with timepoints at 0, 1, 2, 4, and 8 hours (Figure 23). These results reveal that the gamma products from **R16** and **H13** appear immediately, but do not increase in intensity over time, while the **R16** gamma-prime band increases steadily over time (Figure 13). Since the gamma-prime band initially appears in concert with the gamma bands, it may correspond to a similar reaction at a different recombination junction. Another possibility is that the gamma bands are intermediates in the formation of other products; this would explain why the gamma bands do not increase with time but the gamma-prime do.

In the case of **H13**, the alpha-prime band is visible after just one hour, providing still further evidence that this reaction is more robust than its counterpart in **R16**. As with **R16**, the gamma band of **H13** is visible at the zero timepoint, and in the absence of

magnesium, though the presence of magnesium does appear to improve the reaction yield. In addition, the **H13** gamma band is present at time zero with no intermediates, indicative of a fast reaction like the **R16** bands. The overall profile is most similar to the gamma band of **R16**; the **H13** gamma bands do not increase in intensity over time.

A one-hour reaction of radiolabeled **R16** reveals that at least some of the gamma and gamma-prime bands have appeared prior to incubation (Figure 24). For this experiment, **R16** was gel-purified prior to kinase treatment for one hour at 37°C. The first lane is a post-rehydration lane, with the sample taken after organic extraction, ethanol precipitation, and rehydration of the phosphorylated RNA. Since the **R16** was gel purified prior to phosphorylation with ^{32}P, the gamma and gamma-prime bands have either appeared during the kinase step or during the room temperature rehydration in water. The bands are slightly fainter in the post-rehydration step, indicating that the reaction conditions at the first timepoint (immediately after addition of buffer and magnesium) are conducive to the reaction.

Another phenomenon seen in the one-hour incubation of **R16** is an intense cleavage band at about 10-11 nucleotides long. Since this band, like the gamma and gamma-prime bands, appears prior to incubation, it must have formed during phosphorylation or during rehydration of the RNA after ethanol precipitation, and indicates that there is some possible secondary structure of **R16** that creates a labile bond. Another hypothesis is that this is the byproduct of a branching reaction, discussed in the next chapter. The fact that this piece is radiolabeled indicates that the bond fracture must occur towards the 3′ end, possibly at the phosphodiester bond between 10G and 11C or 11C and 12A. In any case, for **R16**, no alpha-prime or beta bands are visible within one

hour. With magnesium held constant at 100 mM and pH 8.0, additional one-hour experiments for both **R16** and **H13** at 37°C, 48°C, and 65°C failed to detect any alpha-prime products by either radioactive labeling or SYBR™ Gold staining (data not shown).

In order to observe whether the concentration of our oligomers had any serious impact on the reactions, we conducted standard self-incubations for both **R16** and **H13** at 10, 20, and 50 μM (Figure 25). These results reveal that the products of both self-reactions readily form over this concentration range. In addition, as a test of whether cold-cycling improved the self-recombination reactions, we compared the aforementioned results to identical self-incubations at constant room temperature (Figure 26). A comparison of the reaction products for **R16** and **H13** at either cold-cycling or isothermal conditions (Figure 27) reveals that cold cycling has substantially improved yield of recombinant products. It can also be observed that the cleavage bands under the cold-cycling regime are reduced in comparison to the isothermal regime.

3.3 Confirmation of RNA activity by transcription

Although the high-throughput sequencing results confirm the presence of recombinant products, we had to consider whether some products, especially the gamma and gamma-prime bands, were artifacts of the synthetic process used to make the RNA. Since the starting RNAs were gel purified prior to reactions, they should be free of contaminants, but it is possible that nucleobases could have been altered during the phosphoramidite synthesis of RNA from TriLink or IDT. Therefore, to confirm that the activity of the RNA was intrinsic to the RNA, we synthesized a modified version of **R16** by runoff transcription with T7 RNA polymerase.

T7 RNA polymerase has absolute requirements for synthesizing short RNA. In particular the polymerase must bind to a promoter attached to the template DNA and the transcript must start with guanosine (cytosine in the template). Although it is possible to substitute adenosine for guanosine for a reduced yield, there is an absolute necessity of a purine at the beginning of the transcript[43]. If this purine is not present, the polymerase may skip to a downstream nucleotide in the template. In addition, the 3′ ends of transcribed RNA can have additional nucleotides that result in a heterogenous mixture and make the correct product difficult to remove by gel purification.

We initially attempted to transcribe **R16** by attaching it to the 3′ end of a hammerhead ribozyme such that subsequent cleavage would release **R16**, but cleavage was relatively poor and the products of hammerhead cleavage appear heterogenous, with bands ranging from 14-17 nucleotides that are difficult to separate (Figure 28). However, we were able to purify some of the presumed 16mer product with a double gel purification (Figure 29). After a 7-day incubation of **R16** as cleaved by the transcribed hammerhead, in our standard conditions, it is clear that the transcript produces the identical products of the synthetic **R16** (Figure 30). In addition, like the behavior of the synthetic **R16**, some of these products are present in the control, likely formed during rehydration. Further work to optimize both the hammerhead, its cleavage, and to purify the product, may be necessary to get absolute confirmation of reactivity.

We next chose to synthesize a modified version of **R16** substituting guanosine for cytosine at the first position of **R16** so that the resulting transcript is 5′-GGUACCGUUGCAUUUG. We also included methoxy additions at the 2′-OH of the first two nucleotides in the template; these modifications have been shown to reduce

heterogeneity in the transcripts[44]. The 5′ modification of **R16** with a G instead of a C may change several aspects of the reaction. In particular, the binding of two complementary 5′ and 3′ ends of **R16** is improved by two full G-C pairs compared to the original **R16**, and the 5′ cytosine, which seems to improve 5′ ligation to a cyclic phosphate in random sequence RNA, is eliminated. However, we reasoned that the 1-G substitution should not completely eliminate all of the products.

Using the DNA template 5′-CAAATGCAACGGTACCTATAGTGAGTCG TATTA, which consists of the 16-nucleotide region attached to the T7 promoter complement, the products of T7 RNA polymerase transcription were treated with DNase I, and subject to organic extraction and ethanol precipitation. We also gel purified the products with UV shadowing using a 16-nt size marker. Despite our use of methoxy modifications in the template, the transcription does produce 3 distinct products. Each of these products was gel purified, ethanol precipitated, and reacted for 7 days of cold cycling using our standard reaction conditions.

As with the **R16** reaction derived from the hammerhead construct, the reactions of the R16-1G products do produce well-defined product bands (Figure 31). Here, it is abundantly clear that the transcribed RNAs in the negative controls show no recombinant products, but after a standard incubation, each one produces recombinant products. The middle one of these bands, which should correspond to the true R16-1G product, has a profile that is very similar to **R16** – there are two general regions of products that correspond to the putative alpha-prime and gamma-prime bands, with several bands in between. In addition, the -1 and +1 transcription products also produce product bands after a standard incubation, with the products shifted downward and upward by a similar

migration factor. Importantly, the generation of products by the R16-1G confirms that the general capacity to recombine is not caused by impurities from the synthetic process.

3.4 Putative structures of recombinant products

The small size of our recombinant products means that it should be relatively easy to predict the secondary structure with computer models. Using the RNA Folding Form on the mFold[45] web server and pKiss from the RNA Shapes Studio[46], we computed the predicted secondary structure of **R16** and **H13** and their respective linear products. Notably, no significant secondary structure is predicted for either **R16** or **H13**. According to mFold, the base **R16** RNA folds into a very weak tetraloop hairpin with a minimum delta-G of -0.1 kcal/mol, while **H13** folds into a similarly weak tetraloop with a minimum delta-G of +0.6 kcal/mol (Figure 32). Thus, it is likely that a large proportion of both of these RNAs are linear at room temperature.

In contrast, the predicted secondary structures for linear recombination products of the **R16** and **H13** self-reactions have considerably lower free energy and likely exist in one or more folded states. According to mFold, for **H13**, the predicted secondary structure of the alpha-prime product at 25 nucleotides is a hexaloop hairpin closed with a five-nucleotide helix and delta-G of -6.70 kcal/mol while pKiss predicts the same 25mer to exist as a pseudoknot with a delta-G of -10.80 kcal/mol (Figure 33).

Each of the linear recombination products of **R16** is predicted to have a stable hairpin by mFold, while pKiss predicts each one to have a pseudoknot as the lowest-energy secondary structure. For the alpha-prime product at 28 nucleotides, mFold predicts a stable hairpin with a delta-G of -3.0 kcal/mol while pKiss predicts a pseudoknot (Figure 34). The alpha-prime product at 29 nucleotides has two possible

hairpins with equivalent delta-G values of -3.5 and -3.6 kcal/mol, with pKiss again predicting a pseudoknot (Figure 35). Similarly, the alpha-prime product at 30 nucleotides is predicted to have a hairpin with delta-G of -5.10 kcal/mol in addition to a possible pseudoknot structure (Figure 36). Finally, the beta product is predicted to be a stable hexaloop hairpin with a delta-G of -6.70 kcal/mol (Figure 37) – the pseudoknot in this case is identical to the hairpin but with part of its 3′ tail hydrogen-bonded to some of the loop nucleotides.

As a final assessment of the structure of our linear recombination products, we attempted to characterize the nature of the linkage formed in the recombination reaction. It has been demonstrated that the alpha reaction of Lutay et al. produces approximately 95% 2′-5′ linkages in the recombinant products. For our own RNAs, whose 50% G-C content makes RNase analysis difficult, we examined the nature of the bond by subjecting the linear recombination products to base-catalyzed hydrolysis. We gel-purified only the linear recombination products, radiolabeled them, and digested them with 1M NaOH at 90°C with timepoints at 0, 60, and 120 seconds (Figure 38).

If the alpha-prime products of R16 and H13 were to contain 2′-5′ linkages, we would expect strong bands corresponding to digestion at the recombination junctions. For **R16**, the recombination junctions for the alpha-prime 28, 29, and 30-nt oligomers would be the 12^{th}, 13^{th}, or 14^{th} bonds in the products, but none of the digestion products at these lengths show any significant intensity. Instead, it appears that the **R16** reaction products show strong bands at the zero timepoint whose lengths are equivalent to the starting material (i.e. **R16**). However, this is not the length that would manifest with a 2′-5′ linkage, which is known to be more labile[47]. The fact that only a radiolabeled 16mer is

present indicates that this is either contamination of the starting material that comigrates with the products, or that the peculiar secondary structure of one or more of the recombinant products produces a labile bond at the junction of the 16th phosphodiester bond. The lack of strong radiolabeled products at the true recombination junction points is an indication of a 3′-5′ linkage at that position. For the **H13** alpha-prime product the result is similar – cleavage of a 2′-5′ bond at the recombination junction should produce a 12-nucleotide oligomer but there is only a strong radiolabeled band at 13 nucleotides, which cannot be the recombination junction of an alpha-prime product. As with **R16**, the quantities of radiolabeled **H13** digestion products at the true recombination junction do not suggest a particularly labile bond.

In summary, we have designed two small oligomers, **R16** and **H13**, that have the ability to undergo self-recombination at slightly basic pH and in the presence of magnesium to produce a range of larger products anywhere from 20 to 43 nucleotides in length, with the bulk of the products being within the range of 28-38 nucleotides, as judged by gel migration and HTS. The yield of the reaction products is small and likely does not exceed a total of 5%, but high throughput sequencing confirms the presence of several linear products, including the most prevalent product at 29 nucleotides and several others from 28 to 31 nucleotides. We hypothesize that the lack of sequencing products above 31 nucleotides, despite strong bands on the gel at 34-36 and 38 nucleotides, is because these are nonlinear RNAs. We have also provided evidence that a reasonable proportion of the linkages formed in the reaction are 3′-5′ linkages, and we have provided additional evidence suggesting that the RNA recombination products likely have considerably more secondary structure than the starting material. In addition,

the countermeasures we employed against contamination, and the runoff transcription of an **R16** analog with a substituted 5′ end, are convincing evidence that the reactivity is intrinsic to the RNA and not a byproduct of the synthetic process of the suppliers. Finally, the generation of recombinant products by the transcribed RNA is proof-of-principle that small RNAs can recombine into longer products.

Chapter 3 Figures

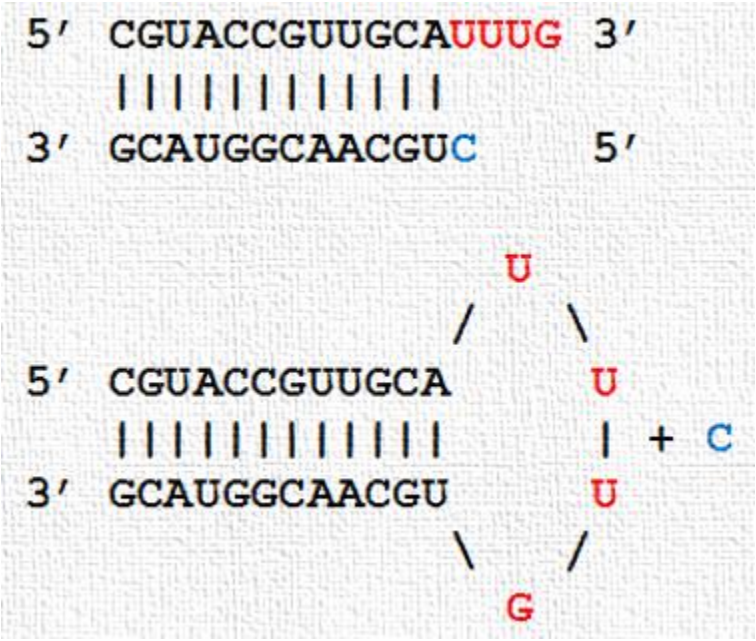

Figure 6: Lowest energy double-strand conformation of R16 and H13 together (top). Attack of the terminal guanosine on the phosphodiester bond of the 5′ cytosine (blue) may produce a hairpin with a cytidine leaving group, similar to the mechanism proposed by Pino et al.[37]

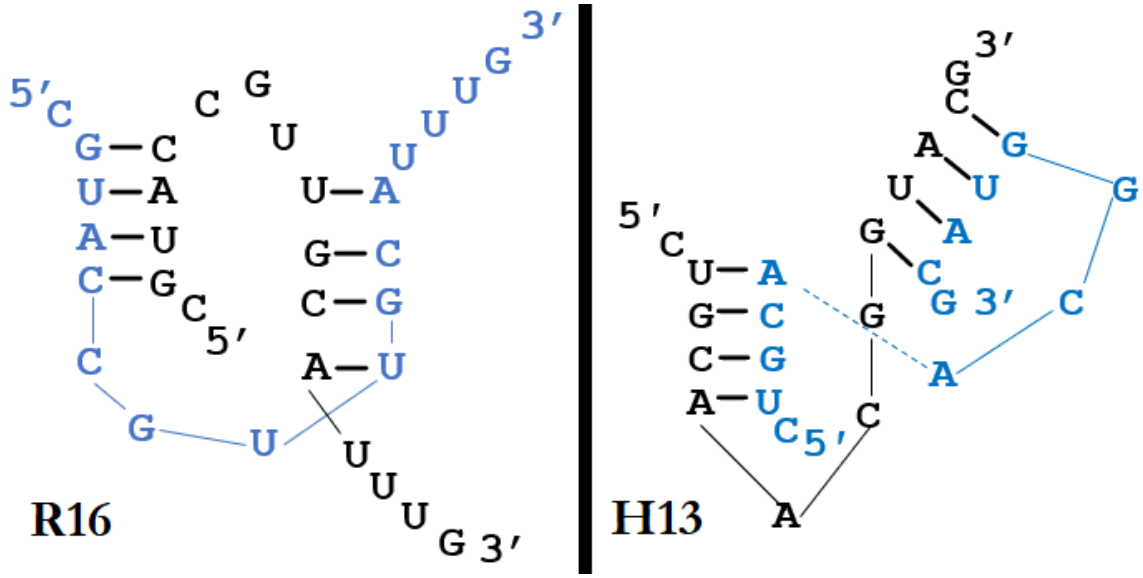

Figure 7a (left): Twisted double strand conformation of R16. 7b (right): Twisted double strand conformation of H13

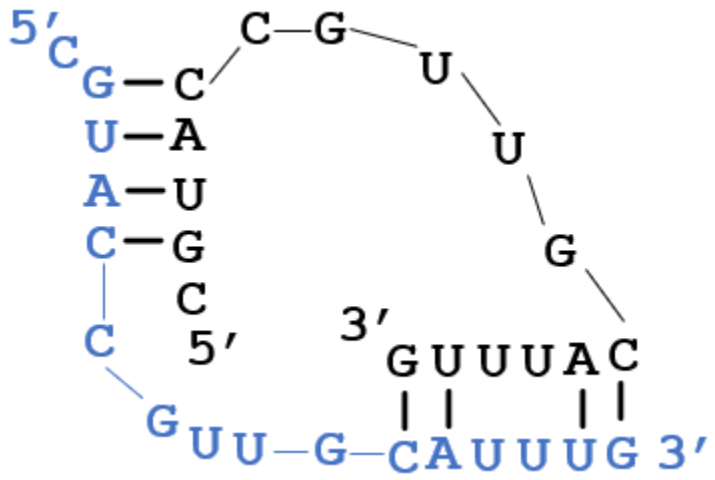

Figure 8: Twisted duplex of R16 with shifted base-pairing and non-canonical U-U pairs.

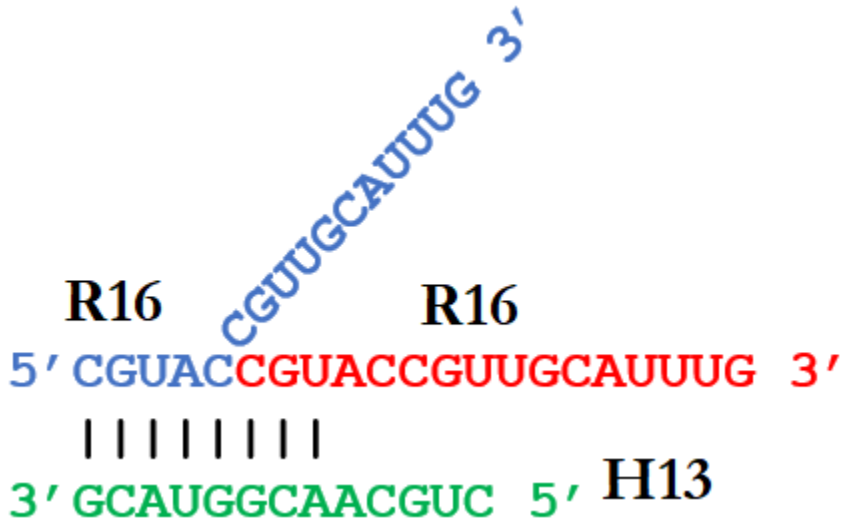

Figure 9: Possible alpha setup for two R16 oligomers splinted over an H13 oligomer. The 11-nt tail of the first R16 may be specifically cleaved analogous to the alpha reaction of Lutay et al. and subsequent ligation of the 5′-OH of the rightmost R16 to the cyclic phosphate would produce a 21-nt product.

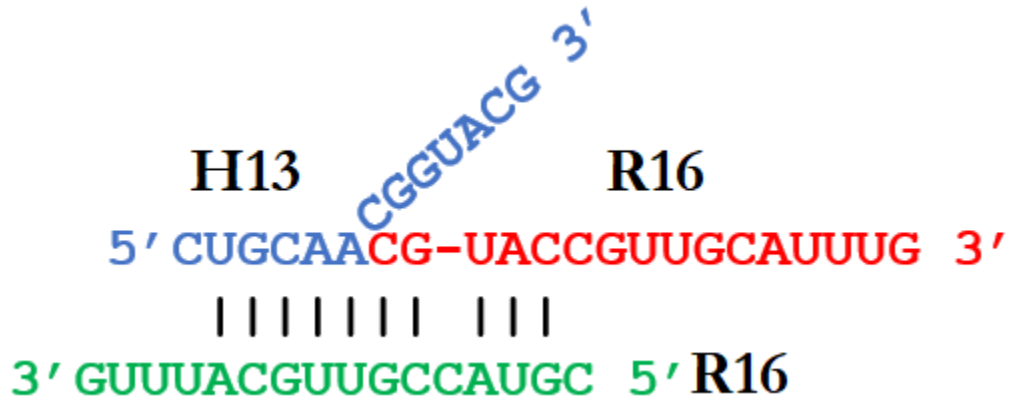

Figure 10: Possible alpha triplex of R16 and H13 with a single-nucleotide bulge in the splint, which would produce a 22-nt product.

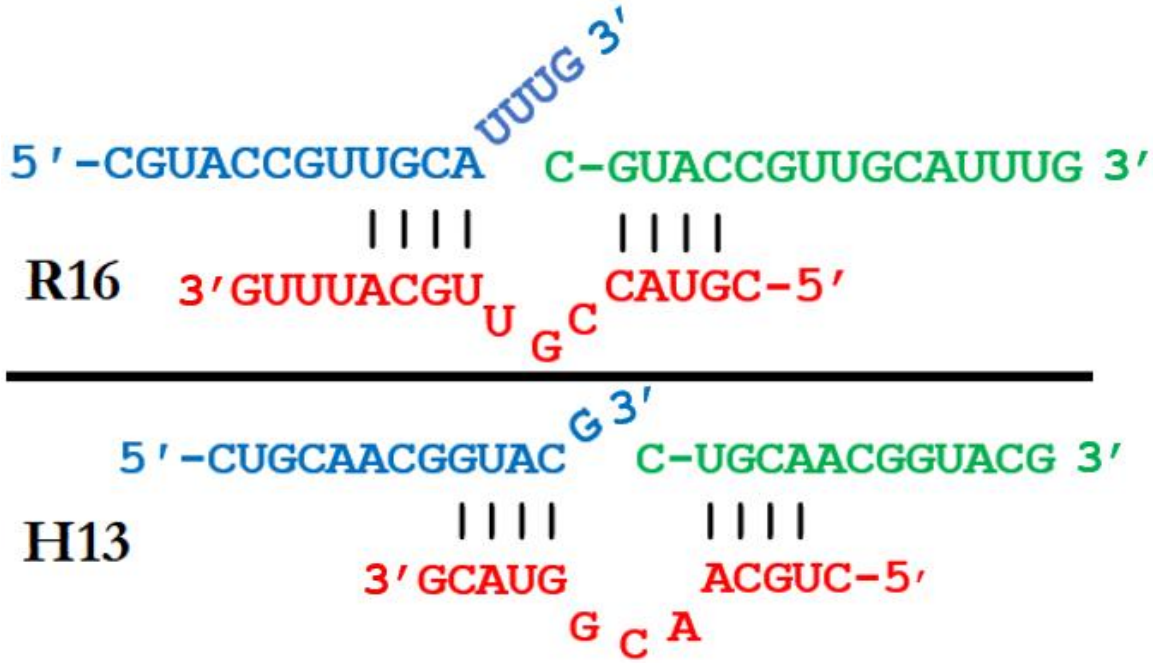

Figure 11: a (top): self-templating triplex of R16 with four Watson-Crick pairs on either side of a 3-nt bulge and an overhanging 3′ -UUUG tail. B (bottom): self-templating triplex of H13 similar to R16 but with single nucleotide -G in 3′ tail. Both complexes have an unbound 5′ cytidine above the bulge.

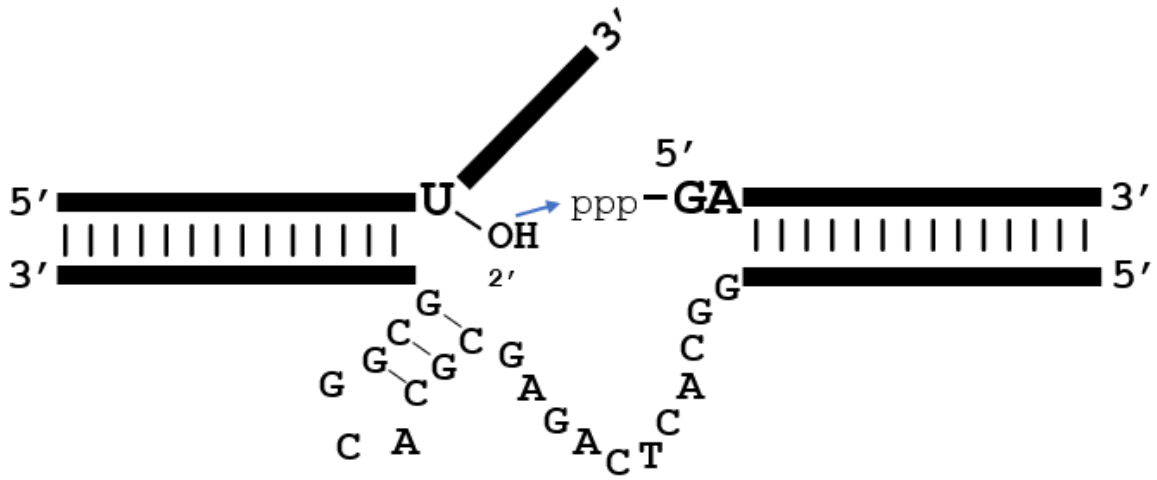

Figure 12: The deoxyribozyme-catalyzed branch reaction of Silverman, featuring two splinted RNAs over a complex bulge. The 2′-OH of a uridine residue in the upper-left substrate attacks a 5′ triphosphate to form a branched RNA. Figure illustrated from reference 41.

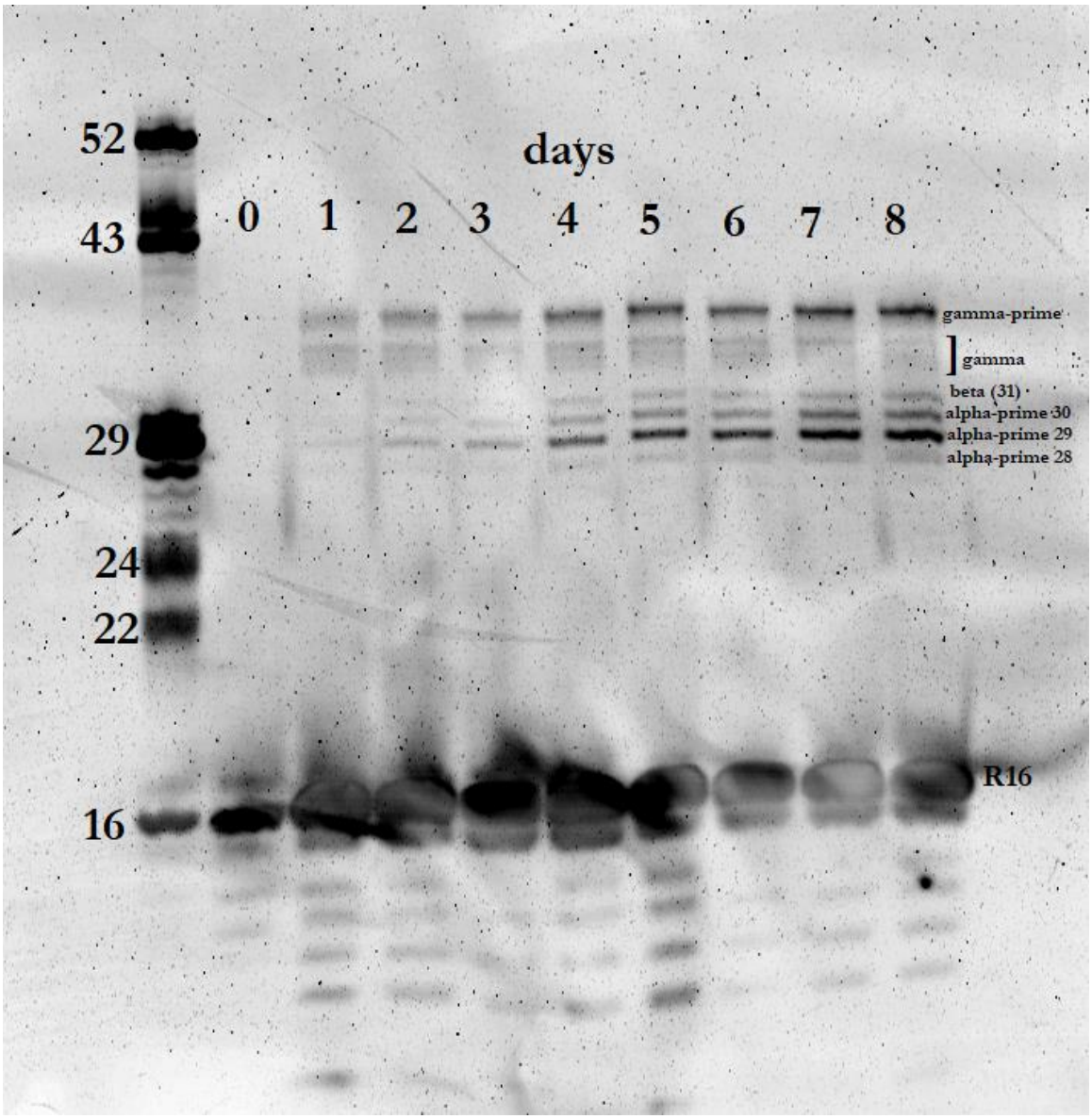

Figure 13: R16 recombination reaction over 8 days with one-day timepoints. The alpha-prime region consists of linear bands from 28-30 nucleotides long, with the strongest band manifesting at 29 nucleotides. Beta is a linear band migrating at 31 nucleotides with a different mechanism. The gamma bands are a region of bands more than twice the size of the starting material at 34-36 nucleotides long, and gamma-prime is one or possibly two bands migrating at about 38 nucleotides, also more than twice the size of the reactants.

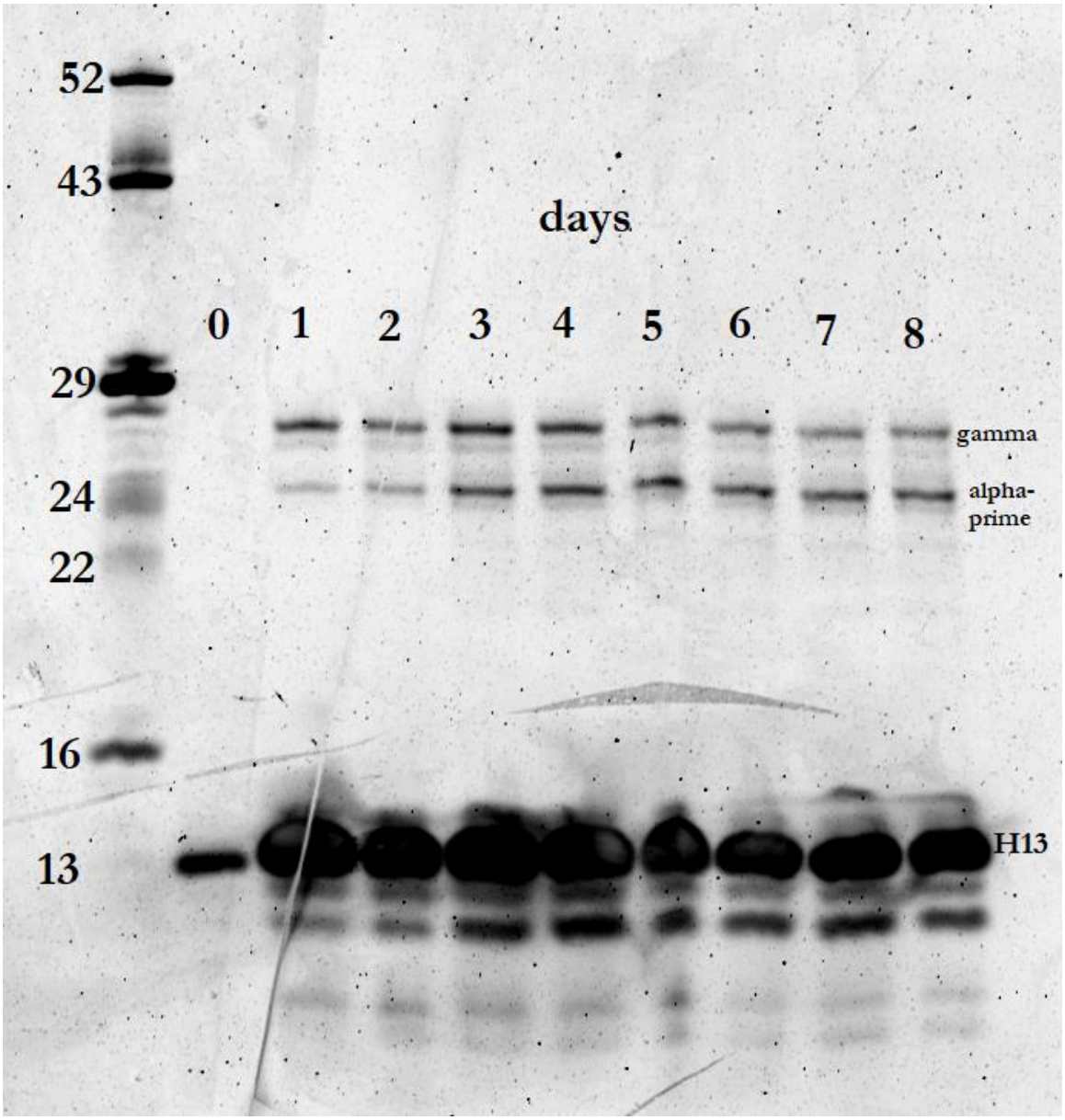

Figure 14: H13 recombination reaction over 8 days with one-day timepoints. There is one gamma region and one alpha-prime region, each of which has a predominant band and a lesser band. For the H13 alpha-prime, the major product is at 25 nucleotides and the lesser is at 24, whereas the main product of the gamma reaction is ~27-28 with a lesser product one nucleotide below it.

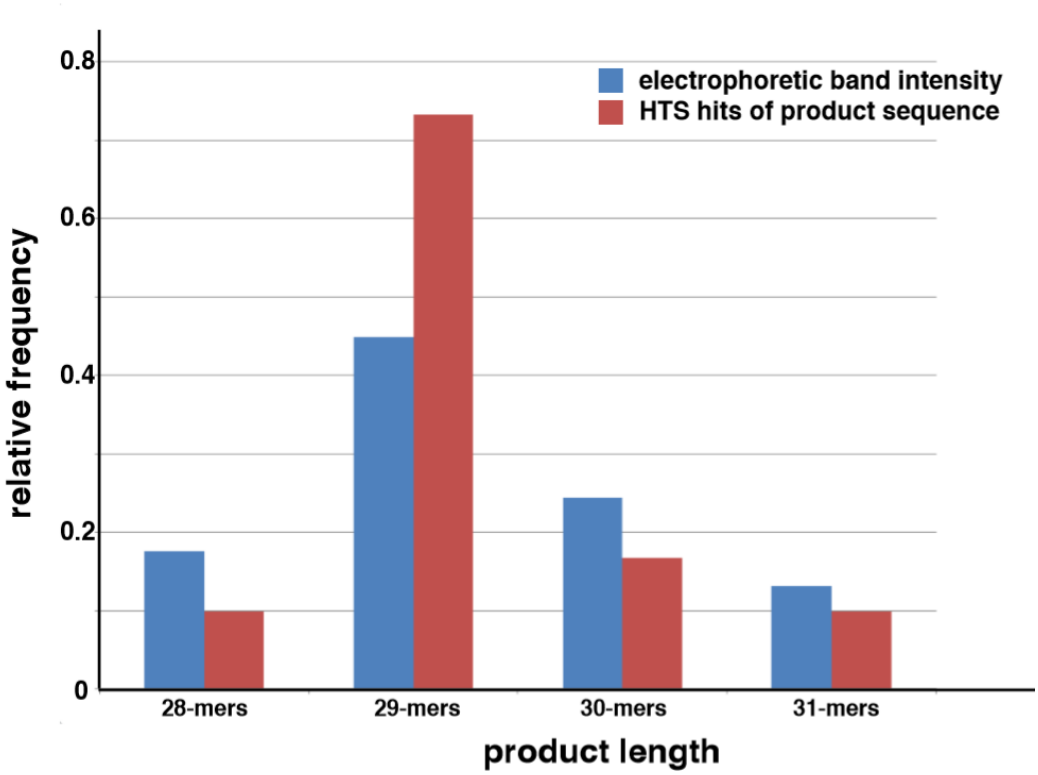

Figure 15: Comparison of relative electrophoretic band intensity and HTS hits of product sequence for linear recombination products. For this graph, the intensities of the linear products were compared to each other, and not to the starting material.

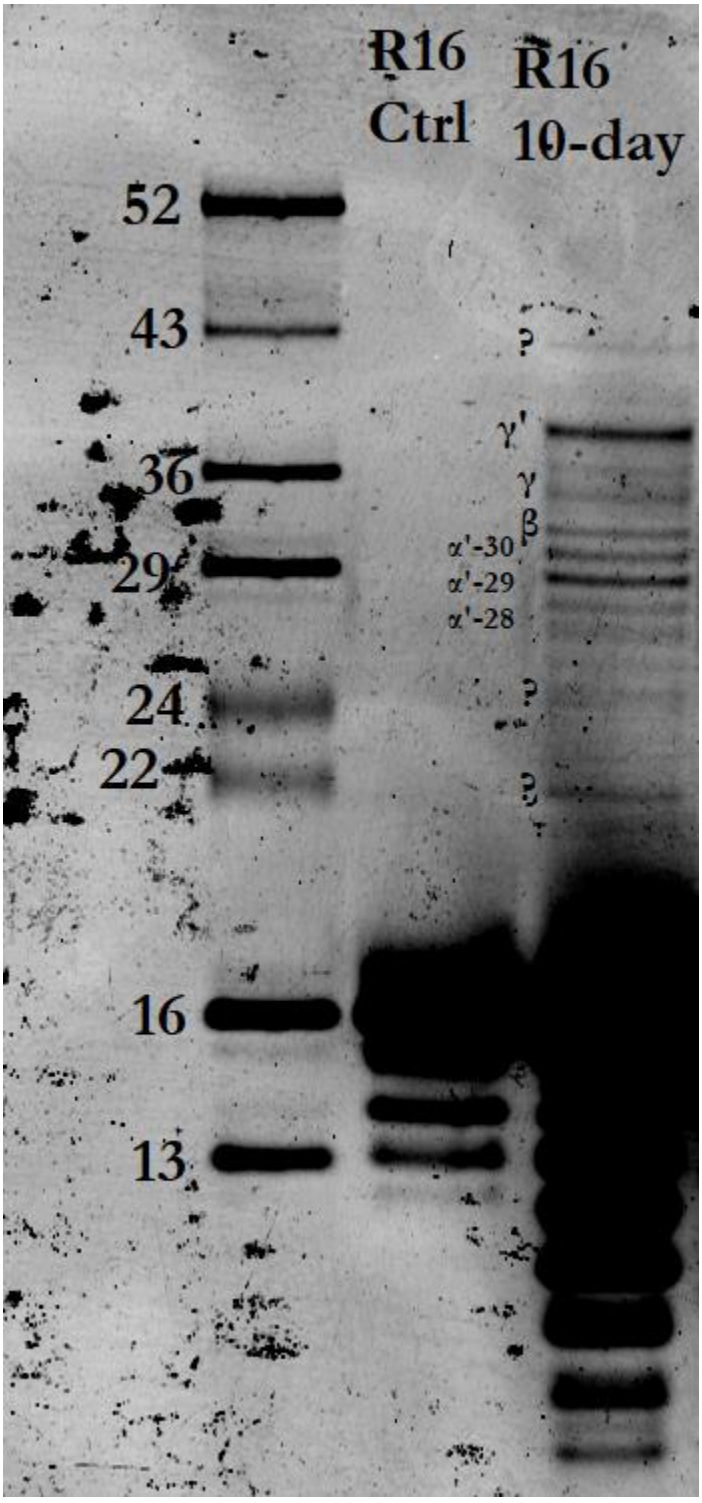

Figure 16: A 10-day reaction of R16, showing additional unknown products near 43, 22, and 25 nucleotides.

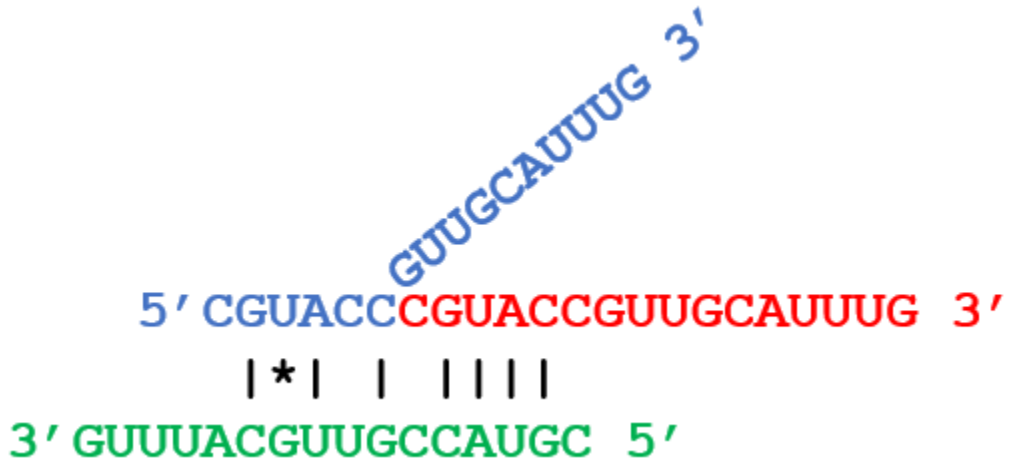

Figure 17: Possible alpha complex of R16 with 22-nt product. This size is detected with visualization of recombination products but is not present in HTS sequencing.

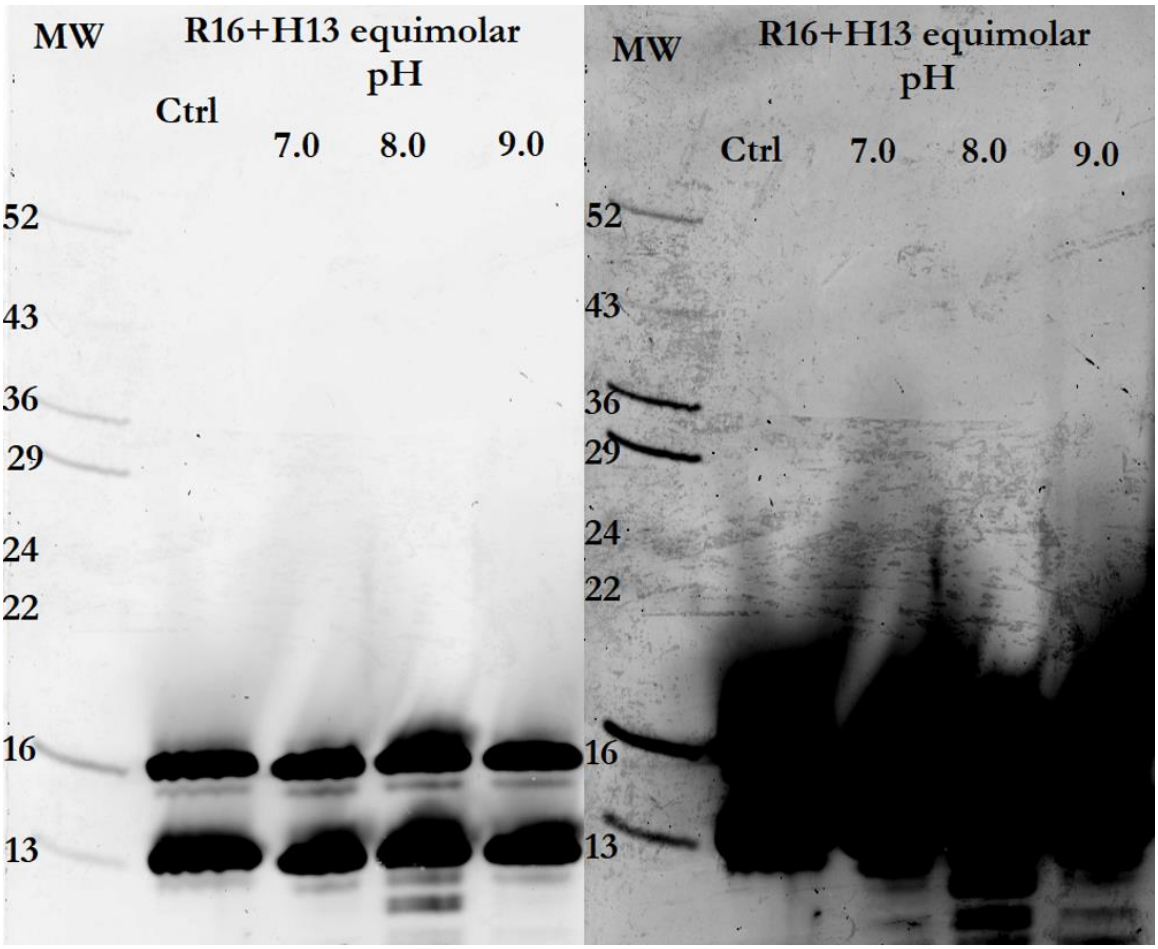

Figure 18: At equimolar concentrations, R16 and H13 mutually inhibit the self-reactions of the other and no products can be seen at a range of pH values. Left: low contrast version of gel showing starting material. Right: High contrast confirms complete absence of products in joint reaction when concentrations are equimolar. For this experiment, RNAs were gel purified prior to incubation and heated together prior to the addition of buffer and magnesium.

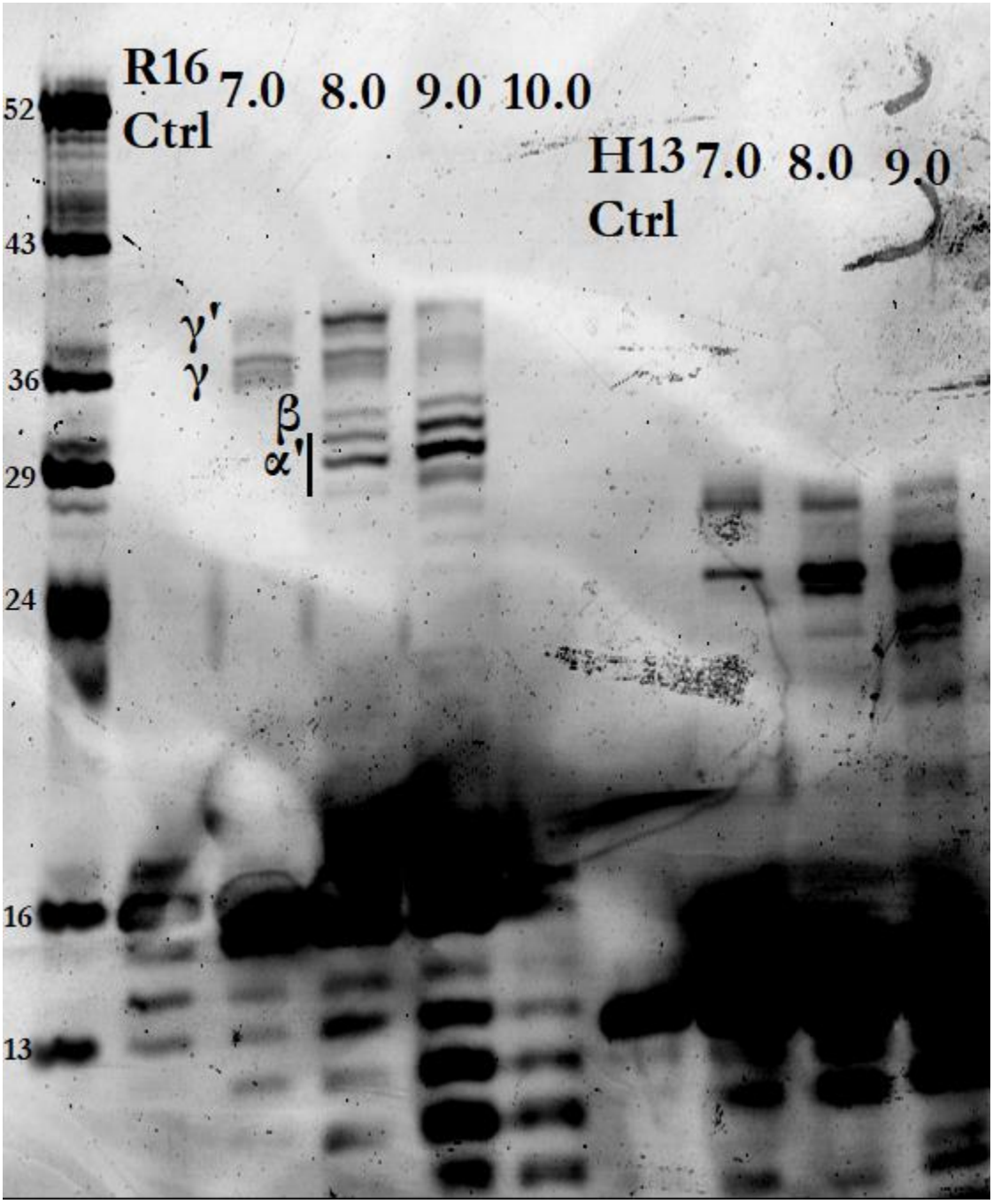

Figure 19: R16 and H13 self-reactions at different pH values.

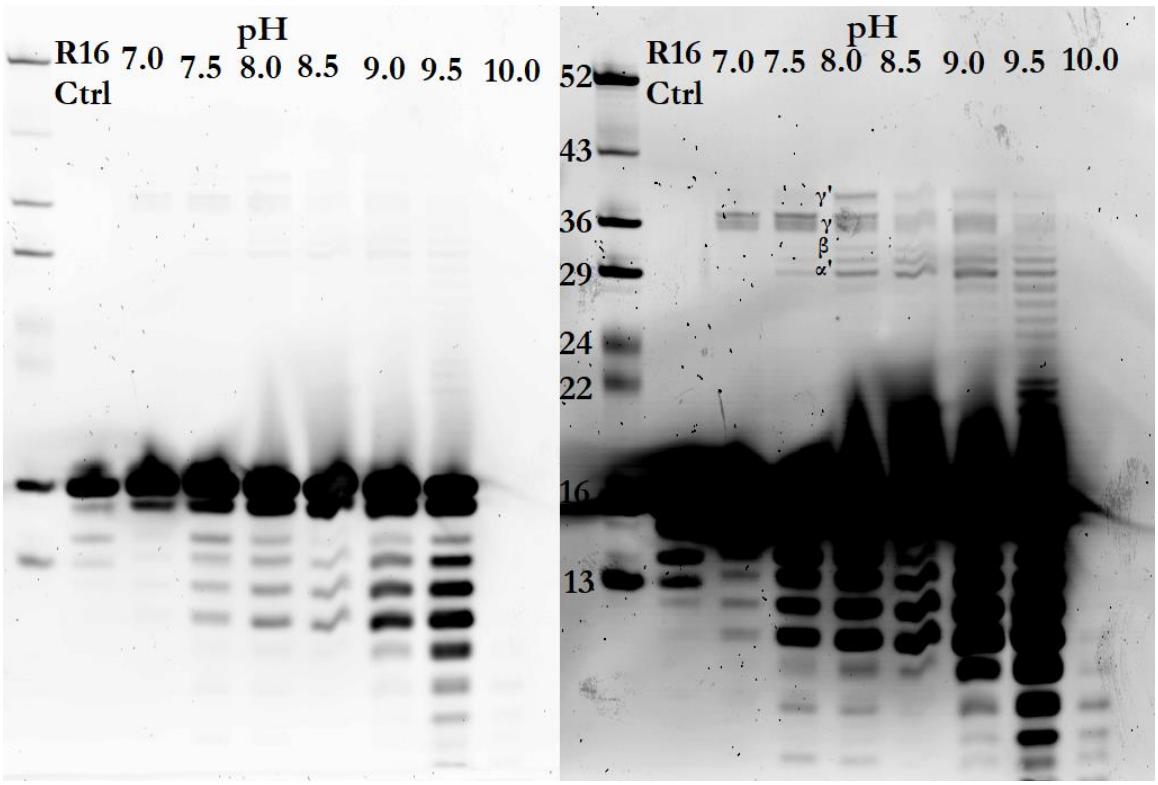

Figure 20: pH profile of R16 recombination reaction. Left: High contrast version showing recombinant products. Right: low contrast version showing starting material.

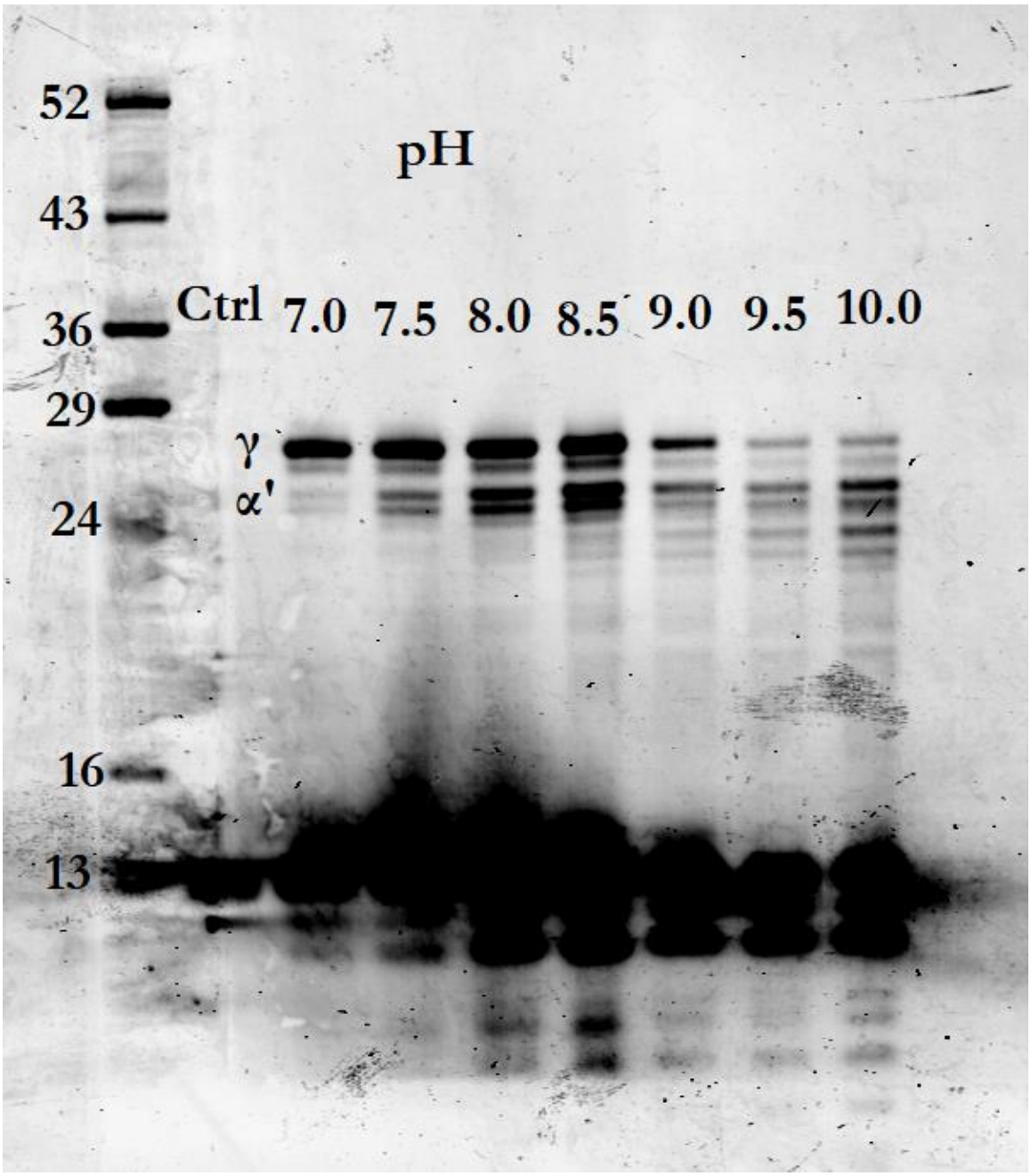

Figure 21: pH profile of H13 reaction

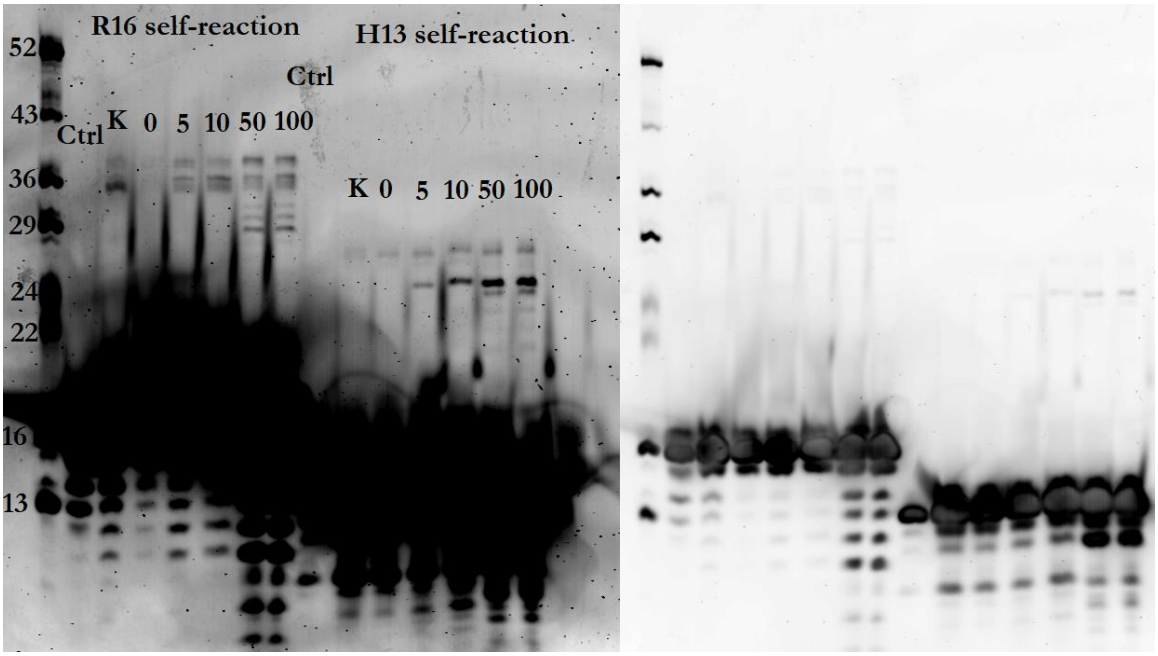

Figure 22: Variance of magnesium concentration for R16 and H13 self-reactions. Left: High contrast version showing products. Right: low contrast version showing starting material. The designation 'K' refers to 100 mM potassium chloride. Numerical values represent mM magnesium chloride. The reaction was carried out for five days of cold cycling with pH constant at 8.0.

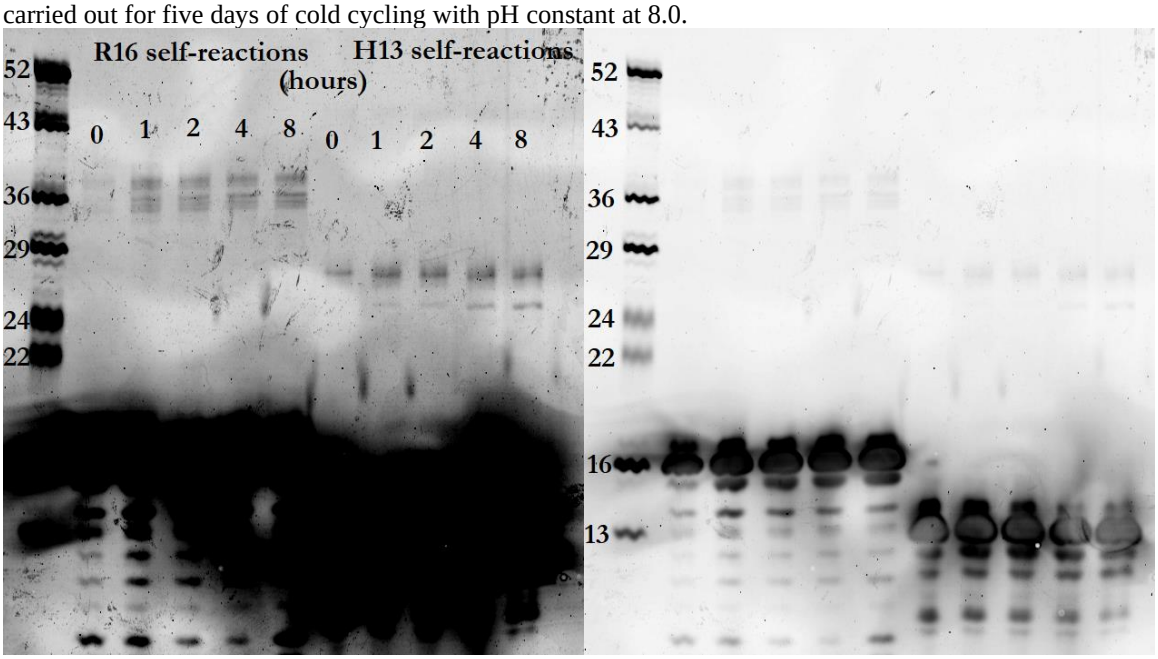

Figure 23: Self-recombination of R16 and H13 with timepoints at 0, 1, 2, 4, and 8 hours. Left: high contrast version showing products. Right: low contrast version showing starting material. RNA was reacted in 50 mM Tris pH 8.0 and 100 mM magnesium at room temperature.

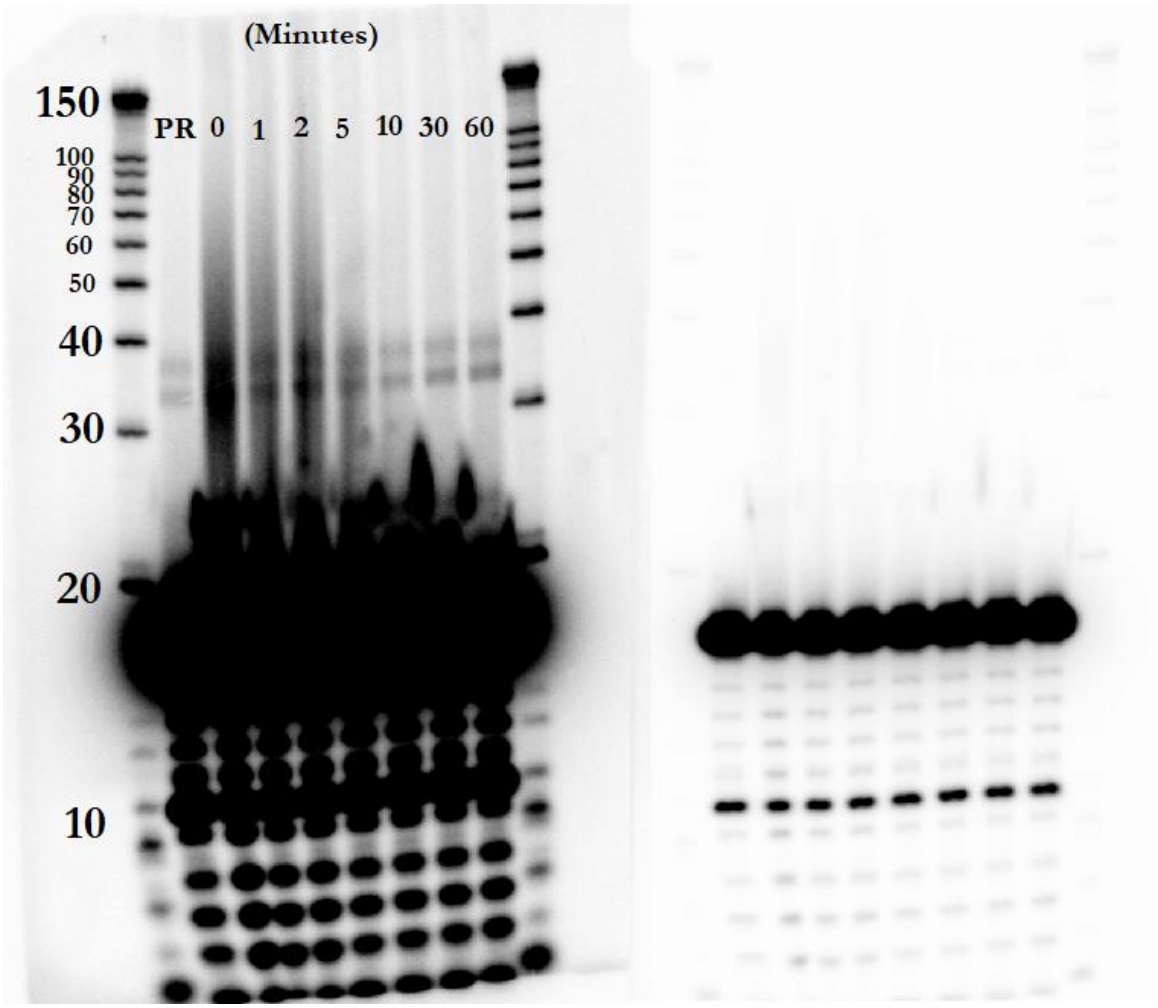

Figure 24: One-hour reaction of R16. Left: high contrast version showing products. RNA was radiolabeled, subject to organic extraction, ethanol precipitated, and rehydrated in 9 µL water. Right: low contrast version showing starting material and cleavage/hydrolysis band at ~10 nt. A one-microliter aliquot was taken after rehydration was complete (PR). Buffer and magnesium were then added to a final concentration of 50 mM Tris pH 8.0 and 100 mM magnesium respectively and a one-microliter aliquot was immediately taken after addition of magnesium for the zero timepoint. Additional one-microliter timepoints were then taken accordingly during room temperature incubation for up to 60 minutes.

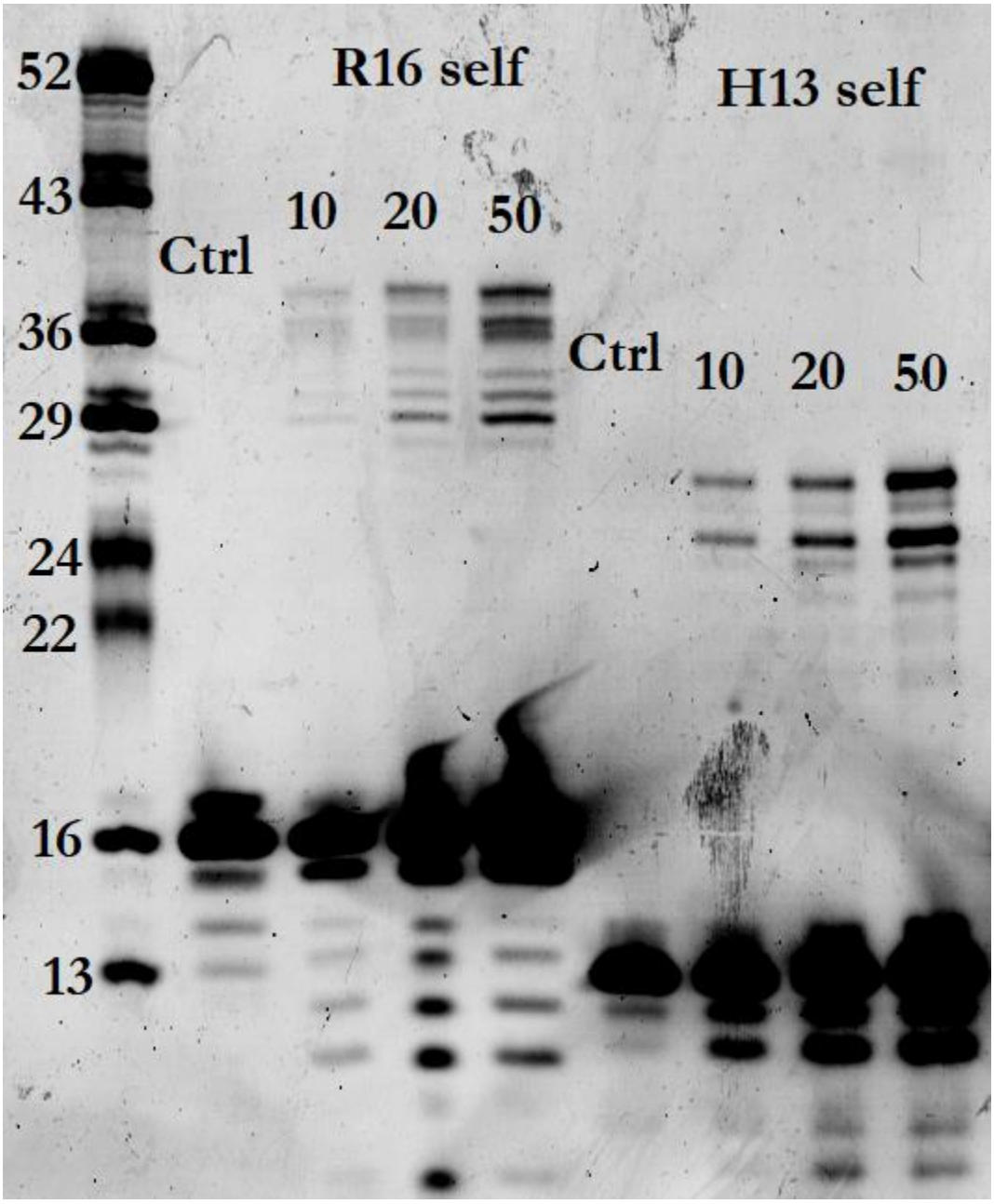

Figure 25: Self-incubations of R16 and H13 at different oligomer concentrations (in µM), under our standard conditions.

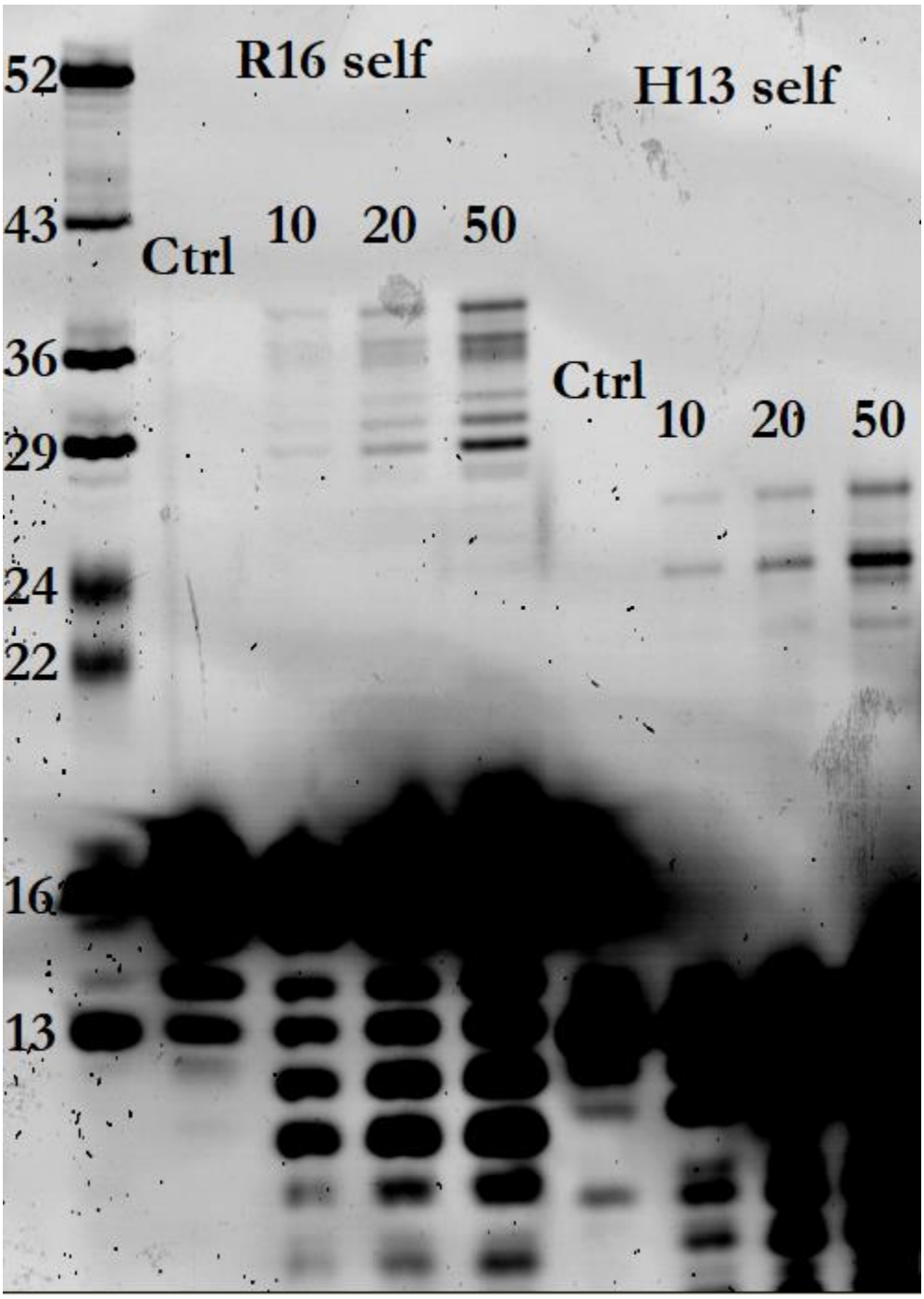

Figure 26: Isothermal concentration variance (in μM) of R16 and H13 self-reactions. These reactions are identical to those in Figure 25 except they were done at constant room temperature instead of cold-cycling, the latter of which was our standard recombination procedure.

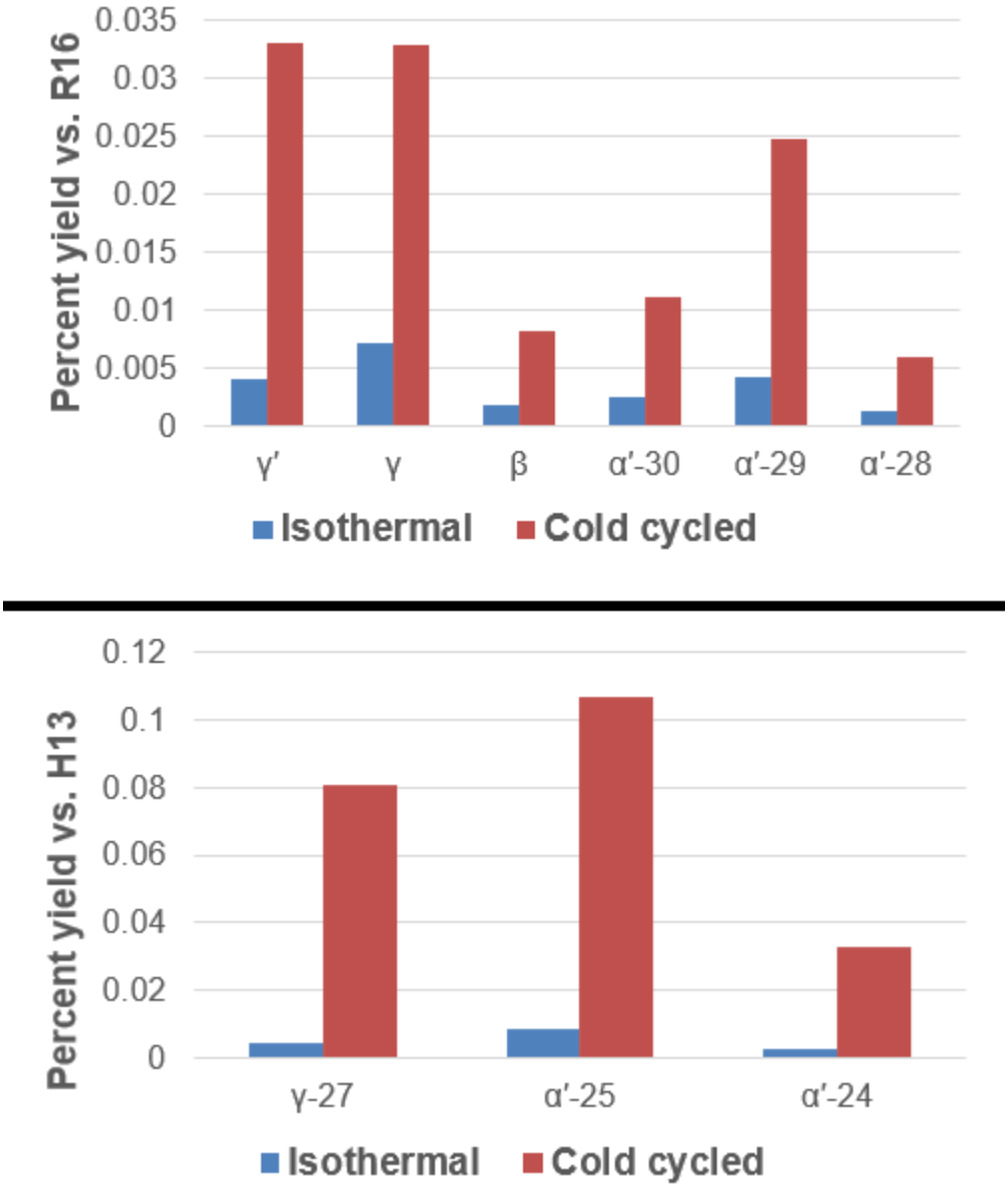

Figure 27: Comparison of intensity between isothermal self-reactions and self-reactions in cold cycles for R16 (above) and H13 (below). Intensities are measured from the 20 μM reactions of Figure 25 or Figure 26 and compared to the band corresponding to the starting material to get the relative percent yield.

Figure 28: Hammerhead cleavage to produce R16. HH+ indicates the full transcript with the hammerhead and R16 while HH- is the cleaved hammerhead.

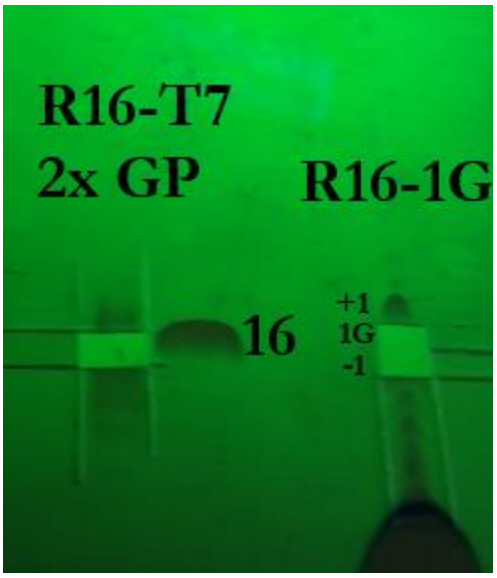

Figure 29: Double gel purification of R16 cleaved hammerhead transcript along with gel purification of three products generated by transcription of the template for R16-1G. The middle product should be R16-1G, 5′-GGUACCGUUGCAUUUG while the others are +1 and -1 products. The middle lane is R16 from IDT.

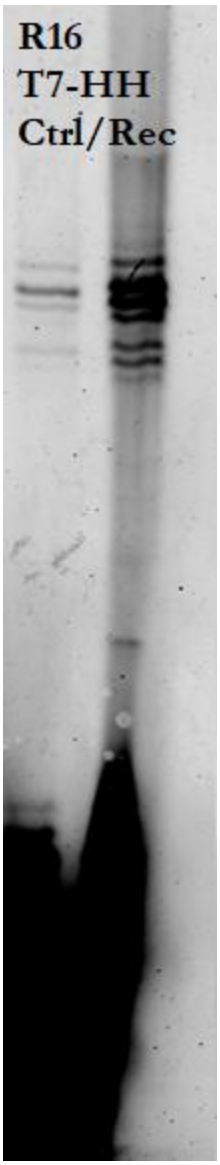

Figure 30: Reaction of R16 produced via run-off transcription of a hammerhead construct. The left lane (Ctrl) is an unincubated control, the right lane (Rec) is a seven-day reaction of RNA in 50 mM Tris pH 8.0 and 100 mM magnesium with 3-hour cold cycles at 22°C and 0°C.

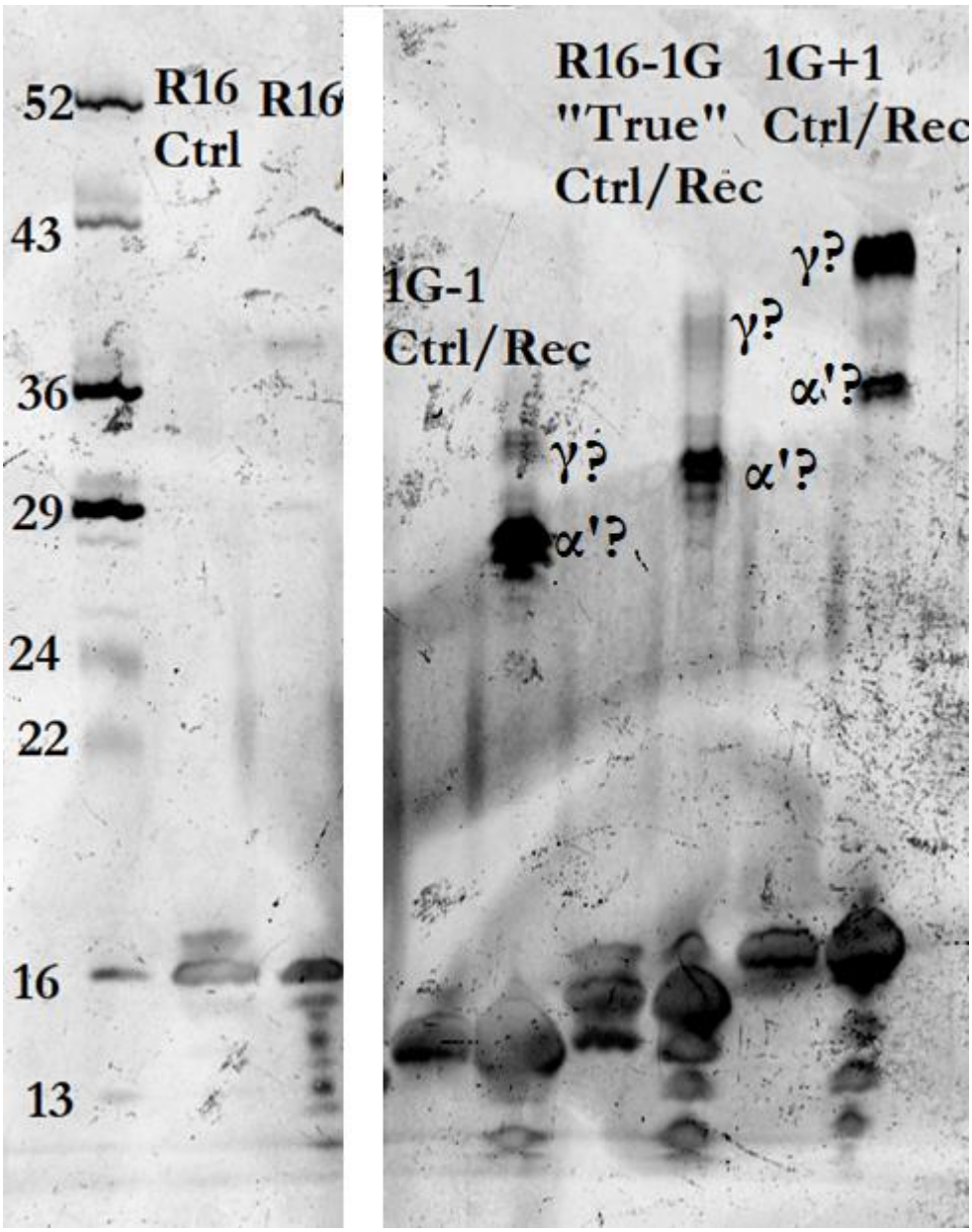

Figure 31: Reaction of transcribed and hammerhead-derived R16 along with three oligomers generated by transcription of R16-1G. R16 from IDT is shown in the first two lanes after the weight markers. All reactions were done for seven days at pH 8.0 in 100 mM magnesium with cold cycling.

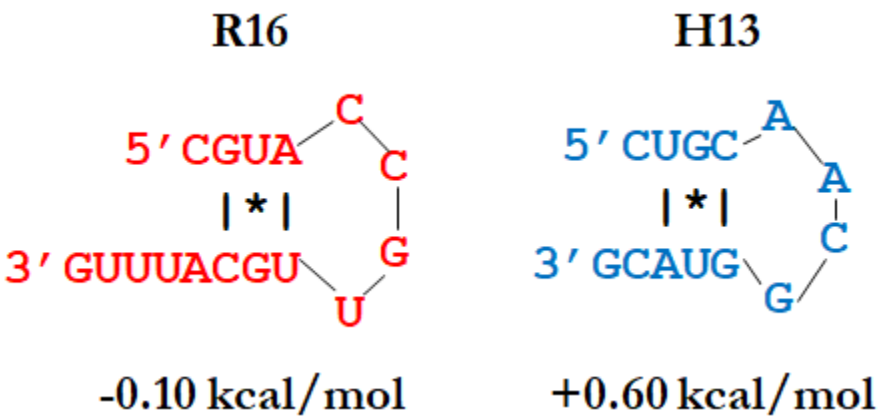

Figure 32: Predicted secondary structures for R16 (left) and H13 (right). Structures were predicted with the RNA Folding Form at the mFold web server (http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form)

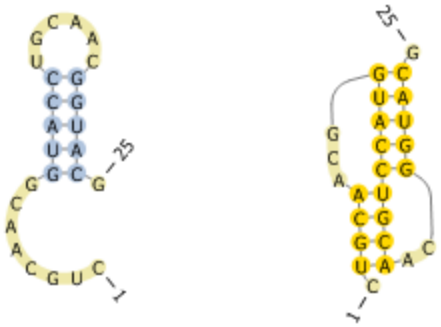

Figure 33: Predicted secondary structures for the alpha-prime recombination product of H13. A stable hairpin (left) and pseudoknot (right) are predicted by mFold and pKiss respectively. Images are illustrated using Pseudoviewer (http://pseudoviewer.inha.ac.kr/)

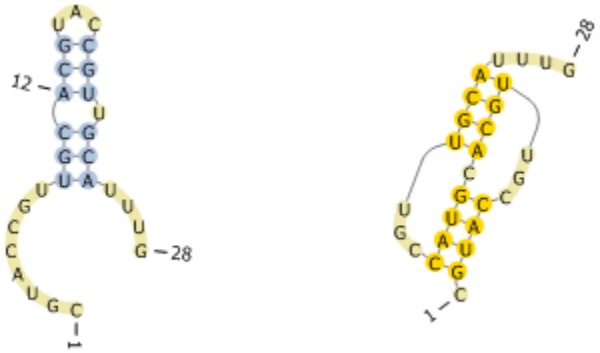

Figure 34: Predicted structures of the alpha-prime 28-nt product from R16 recombination. Images illustrated using Pseudoviewer (http://pseudoviewer.inha.ac.kr/)

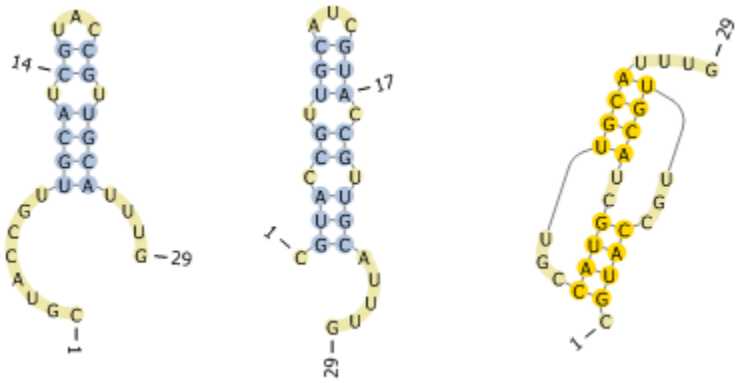

Figure 35: Predicted structures of the alpha-prime 29-nt product from R16 recombination. Images illustrated using Pseudoviewer (http://pseudoviewer.inha.ac.kr/)

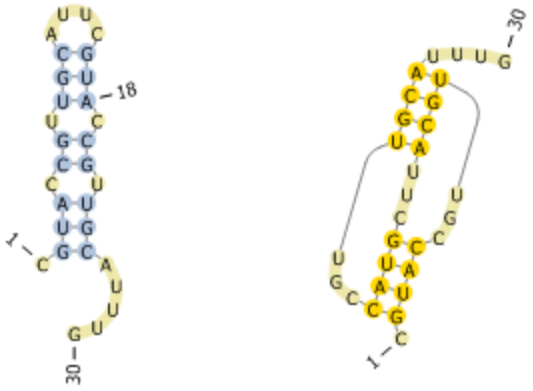

Figure 36: Predicted structures of the alpha-prime 30-nt product from R16 recombination. Images illustrated using Pseudoviewer (http://pseudoviewer.inha.ac.kr/)

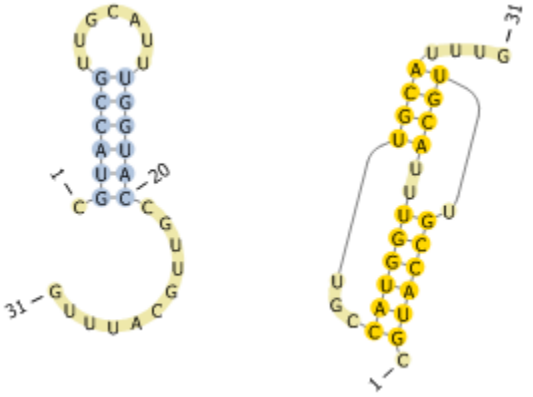

Figure 37: Predicted structures of the beta product from R16 recombination. Images illustrated using Pseudoviewer (http://pseudoviewer.inha.ac.kr/)

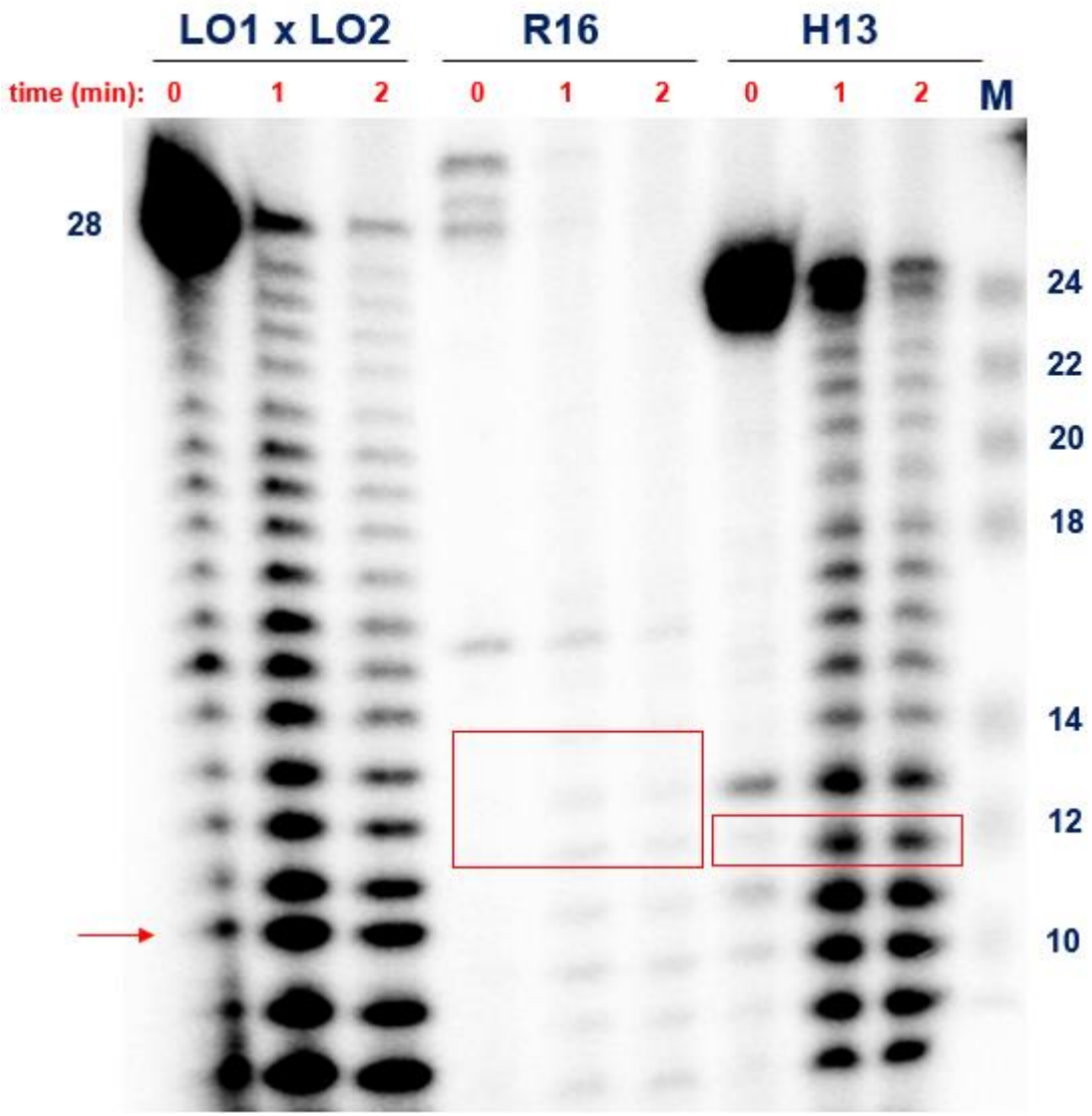

Figure 38: Radiolabeled basic digestion of R16 and H13 products as compared to the Lutay et al. recombination products (LO1 X LO2) and the molecular weight marker (M). The red boxes represent expected digestion products at the recombination junctions. The red arrow identifies the 12mer that results from cleavage of the 2′-5′ linkage in the Lutay et al.[38] recombination product.

Tables

TABLE 1. HTS analyses of **R16** self-recombination products; 20–38 nt gel region after 7 d

Sequence (5′ to 3′)	nt	N	Notes
R16 self-recombination (total counts = 21,746)			
{cgu acc guu gca uuu g}–1	15	331	R16-like 15-mers
–GU ACC GUU GCA UUU G	15	2916	R16 Δ5′C (γ′)
CGU AC– GUU GCA UUU G	15	42	R16 ΔC6 (γ′)
CGU ACC GUU GCA UUU G	16	19437	R16 per se (γ or γ′)
CGU ACC AUU GCA UUU G	16	24	R16 G7A (γ or γ′)
CGU ACC UUU GCA UUU G	16	79	R16 G7U (γ or γ′)
CGU ACu GUU GCA UUU G	16	13	R16 C6U
CGU ACC cGU UGC AUU UG	17	316	R16 + C (γ or γ′)
CGU ACC GUU GgC AUU UG	17	251	R16 + G (γ or γ′)
CGU ACC GUU GcC AUU UG	17	65	R16 + C (γ or γ′)
CGU AC [CGU UGC AU] CGU ACC GUU	22	23	α or α′ recycling
CGU AC [CGU UGC AUU U] GCA UUU G	22	32	α or α′ recycling
{cgu acc guu gca uuu g}	23	14	R16-like 23-mers
CGU ACC GUU GCA U•CG UAC CGU UGC	24	47	α (R16Δ + R16Δ)
{cgu acc guu gca uuu g}	25	22	R16-like 25-mers
CGU AC [CGU UGC AUU U] CGU AC [CGU UGC]	26	8	α or α′ recycling
{cgu acc guu gca uuu g}	26	3	R16-like 26-mers
{cguaccguugcauuug}	27	4	R16-like 27-mers
CGU ACC GUU GCA CGU ACC GUU GCA UUU G	28	17	α′ at ③
–GU ACC GUU GCA UCG UAC CGU UGC AUU UG	28	3	α′ at ②, then γ*
CGU ACC GUU GCA U•G UAC CGU UGC AUU UG	28	1	α′ at ②, then γ*
CGU ACC GUU GCA UCG UAC CGU UGC AUU UG	29	102	α′ at ②
{cgu acc guu gca uuu g}	29	14	R16-like 29-mers
CGU ACC GUU GCA UUC GUA CCG UUG CAU UUG	30	24	α′ at ①
CGU ACC GUU GCA UUG UAc CCG UUG CAU UUG	30	3	α′ at ②, then γ*
{cgu acc guu gca uuu g}	30	16	R16-like 30-mers
CGU ACC GUU GCA UUU GGU ACC GUU GCA UUU G	31	13	β
CGU ACC GUU GCA UUU GGU ACC AUU GCA UUU G	31	1	β with G22A
CGU ACC GUU GCA UUU GGU ACC GUU UCA UUU G	31	1	β with G25U
CGU ACC GUU GCA UUU G•C GUA CCG UUG CAU UUG	32	3	R16 + R16 ligation?

γ* represents a γ product that creates a branch in which a mutation occurs during reverse transcription during HTS preparation.

Table 1: Sequencing results of R16 recombination products. Table from (24).

TABLE 2. HTS analyses of **H13** self-recombination products; 20–38 nt gel region after 7 d

Sequence (5′ to 3′)	nt	N	Notes
H13 self-recombination (total counts = 10,204)			
CUG CAA C–G UAC G	12	12	H13 ΔG8 (γ)
CUG CAA CGG UAC G	13	9074	H13 per se (γ)
CUG CAA CGG UAU G	13	242	H13 C12U (γ)
GUG CAA CGG UAC G	13	128	H13 C1G
{cug caa cgg uac g}	13	238	H13 single mutants
{cug caa cgg uac g}+1	14	37	H13 single inserts
CUG CAA CGg GUA CG	14	28	H13 + G (γ)
CUG CAA CG [UGC AAC GGU ACG]	20	14	α or α′ recycling
CUG CAA CGC UGC AAC GGU ACG	21	26	α or α′ recycling
CUG CAA CGG UAC UGC AAC GGU ACG	24	141	α′ at ②
CUG CAA CGG UAC CUG CAA CGG UAC G	25	258	α′ at ①
CUG CAA CGG UAC G•CU GCA ACG GUA CG	26	6	H13 + H13 ligation?

γ* represents a γ product that creates a branch in which a mutation occurs during reverse transcription during HTS preparation.

Table 2: Sequencing results of H13 recombination products. Table from (24).

Chapter 4: Mechanistic investigations

Based on the results of experiments described in chapter 3, which included the use of RNA from different vendors (TLB and IDT) and RNA generated from runoff transcription with either a hammerhead ribozyme or a substituted 5′ end, it is clear that the self-reactions of **R16** and **H13** are intrinsic to the RNA and not the result of contamination from the synthetic process. That leads to the important question of what are the mechanisms by which these small RNAs recombine?

There are two potential mechanisms for the linear products of **R16** and **H13** self-recombination that are relevant for these systems. One is the two-step mechanism of alpha recombination; cleavage to give a 2′-3′ cyclic phosphate followed by ligation over a splint. The presumptive alternative is a one-step S_N2 attack of a 5′-OH on an existing phosphor-ester bond, although this has not been reported in the literature. However, because the cleavage step of **R16** and/or **H13** must take place over a bulge, it is not clear that such cleavage would be favorable compared to the alpha setup of Lutay et al.[38], where there are 11 base pairs to form an A-form alpha helix and the overhanging poly-A tail is presumably both displaced and extruded by the second complementary strand.

4.1 Deoxy substitutions of R16 and H13 tails

One approach to assess the possibility of two-step cleavage and ligation vs. one-step transesterification is the use of deoxy residues at key positions in **R16** and **H13**. If the internal 2′-hydroxyls responsible for the cleavage reaction are absent, the two-step reaction should be completely inhibited as the RNA would not be able to self-cleave and form the cyclic phosphate intermediate necessary for subsequent ligation.

Therefore, we chose to address the question of one-step versus two-step cleavage for the linear recombinant products by using modified versions of **R16** and **H13** with 2′ deoxy residues at key positions. For **R16**, we chose to substitute four of the tail nucleotides (12A, 13U, 14U, and 15U) with 2′ deoxy residues. The tail is the overhanging 3′ end of the upper left strand in the self-templating triplex of **R16** (Figure 11); our HTS sequencing data indicates that these phosphodiester linkages are broken and reformed for the alpha-prime products. With this fourfold substituted tail, none of the alpha-prime products are formed (Figure 39), but the gamma products do form, providing two significant findings. Firstly, the alpha-prime bands are likely the result of a two-step reaction with cleavage over the 3-nt bulge followed by ligation of the 5′-OH to the (presumptive) cyclic phosphate left behind. Here, there is the assumption that use of deoxy residues at these four positions would not impact a putative one-step reaction. This is probably a safe assumption[48], but it cannot be ruled out with certainty that deoxy substitution cannot affect a putative one-step reaction. For example, if the 2′-OH coordinated a magnesium ion essential for a one-step reaction, this could also be the reason for the loss of this product. Secondly, this demonstrates that the branching reaction of Silverman[41, 42] does not occur with the 3-nt bulge. I will later provide evidence that the gamma products, are likely the result of a hitherto-unknown branching mechanism.

For **R16**, use of terminal deoxyG produces a distinct, visual downward shift, even as the general profile of the reaction is the same (Figure 39). Although it might appear that the shift is almost a full nucleotide, this is likely because the migration rate of the modified oligomers is slightly different. Notably, the presence of the beta band at the 31-

nucleotide mark is consistent with the beta reaction resulting from attack of the terminal 3′-OH, and not the terminal 2′-OH.

In the case of **H13**, removal of the penultimate 2′-hydroxyl completely eliminates the 25mer band (Figure 40), suggesting that its formation results from a two-step reaction. As with the modified tail of **R16**, an important assumption of this approach is that the loss of a single oxygen atom would not disrupt the reaction in any other way. Use of a modified **H13** oligomer containing a terminal 2′-deoxy residue does not change the reaction at all (Figure 40) other than a slight migration shift, similar to the case with the modified **R16** with a terminal deoxy.

As a baseline, the fact that the chosen 2′-deoxy residues eliminated the alpha-prime bands suggested that the best explanation for the alpha-prime products is a two-step mechanism. The appearance of cleavage bands below 16-nt before the appearance of the alpha-prime bands supports, but does not necessarily enhance this theory; presumably cleavage bands of such sizes could appear regardless of whether the actual reaction was one or two steps.

4.2 Phosphorylation of RNA

The key to identifying the recombination mechanisms is identifying the nucleophile in the reactions. Presumably, the nucleophile in the reactions producing alpha-prime products for **R16** is the 5′ hydroxyl of the 5′ cytidine because the sequencing results show an addition to the 5′ end of **R16** (e.g. CGUACCGUUGCAU**CGUACCGUUGCAUUUG**). We therefore performed a phosphorylation of **R16** and **H13** and reacted the purified, phosphorylated RNA under our standard reaction conditions. The results (Figure 39, Figure 40) show that the alpha-

prime bands are absent, providing strong evidence that the 5′ hydroxyl is indeed the nucleophile for the alpha-prime reactions.

Notably, 5′ phosphorylation does not eliminate the gamma bands in either of the self-reactions so there must be one or more other nucleophiles involved in the formation of those products. As the **R16** construct with four deoxy residues in its tail produces the gamma bands, the nucleophile is not a 2′ hydroxyl from those residues, as one might predict from the Silverman reactions[41, 42]. One potential explanation for the gamma and gamma-prime bands is that an internal 2′ hydroxyl is attacking a phosphodiester bond either *cis* or in *trans* to form a lariat or a branched RNA respectively. As described in the earlier chapter, possible secondary structures of **R16** dimerization include two bent *trans* conformations in which the strands must twist sharply to achieve proper polarity (Figure 7, 8). If these *trans* structures were covalently linked they would form a pseudoknot; in fact, a pseudoknot is the lowest energy structure for all of the linear recombinant products as predicted by the pseudoknot-folding algorithm pKiss (Figures 33-37).

4.3 Internal deoxy substitutions

If the bent secondary structures of R16 could form, it would appear that there could be a sharp bend at either 6-C, 7-G, 8-U, or possibly 10-G if base-pairing is shifted downstream by two nucleotides (Figure 8). We tested a version of **R16** with a deoxy residue at 10G, but the reaction products appear unchanged from **R16** (Figure 41). We then performed reactions with deoxy constructs at 7G and 8U (Figure 42), and a further construct with deoxyriboses at 6C, 7G, and 8U, which correspond to the three nucleotides in the splint region of the self-templating triplex (Figure 43). None of these modified versions eliminates either the gamma or gamma-prime bands, although with the triple-

deoxy substitution there does appear to be a downward band shift for the gamma bands, a possible indication that one of these bands is missing entirely. This result implies that the gamma bands are the result of multiple nucleophiles; the elimination of one nucleophile (e.g. d7G or d8U) does not affect the reaction, but using deoxyribose nucleotides at all three positions eliminates a band. For example, if different nucleophiles were attacking the same junction, the size and molecular weight of the products would be virtually identical and the loss of one nucleophile would not be enough to alter the appearance of the product band.

We also considered whether the nucleophile is from a terminal base-paired nucleotide such as 5C, 9U, or 11C, which are adjacent to an unpaired region in the self-templating triplex or duplex. We tested **R16** analogs with deoxy substitutions at 1C, 5C, 9U, and 11C and examined the recombination products after gel purification and a 7-day reaction (Figure 44). It is evident that none of these deoxy substitution analogs show products different from those of **R16**, although as with the general **R16** profile, it is very challenging to separate the distinct gamma bands on the gel.

In total, we tested modified versions of R16 with combined deoxy residues at 13 of the 16 different positions, 1, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16, but were unable to identify a substitution position that eliminated the gamma-prime band with any of these constructs, despite the fact that positions 5 through 11 are the sites of insertions, deletions, and mutations in the high-throughput sequencing results.

4.4 DNA version of R16

Since identifying the nucleophile(s) for the **R16** gamma and gamma-prime bands did not emerge from experiments with 2′ deoxy analogs, we procured the 16-nt **R16**

DNA (5′-CGTACCGTTGCATTTG) from IDT, gel purified it, and subjected the purified DNA to a seven-day incubation at pH 8.0 in 100 mM $MgCl_2$ with cold cycling. As predicted from the lack of 2′ hydroxyls, no alpha-prime bands are present. Remarkably though, a single band in the region of the gamma and gamma-prime bands (34-38 nt) appears for the DNA reaction as well, and since the DNA migrates about 2-3 nucleotides faster, we conclude this band resembles the gamma-prime band (Figure 45).

Since there are no intermediates present on the gel, it is unlikely that this band is the result of multiple recombination mechanisms. Furthermore, as the DNA has no internal hydroxyls to function as nucleophiles, it's possible that this band is an unnatural branch formed with a phosphoramidate linkage, such as might be formed by attack of a nucleophilic amine group from a purine ring on a phosphodiester bond. An example of a similar phenomenon is a report by Burke claiming phosphorylation of a guanine nucleobase by a ribozyme[49].

In light of our gathered experimental evidence, it is difficult to discern the true nature of the gamma-prime bands. Nonetheless, the absence of the gamma region in the all-DNA **R16** construct does support a mechanism involving internal 2′ hydroxyls of the RNA for the gamma region. The fact that most of the single-nucleotide deoxyribose substitutions do not substantially alter the profile of the gamma region also suggests that multiple 2′ hydroxyls are involved in formation of the gamma bands.

4.5 RNase digestions

To investigate the possibility of branched, lariat, or circular RNAs in the gamma and gamma-prime products, we carried out RNase R digestions of the recombinant products of **R16**. RNase R is a 3′-5′ exonuclease that effectively digests all linear RNAs

but cannot digest circular RNAs or the 2′-5′ linkage of a lariat RNA, although it has been suggested that RNase R can digest Y-branch RNA[50]. We performed a one-hour self-reaction of radiolabeled **R16** at room temperature followed by a deep freeze, and then a further one hour to generate only gamma or gamma-prime products. After a one-hour digestion by RNase R, there is no evidence of the two gamma bands visible in the control (Figure 46). As canonical lariats and RNA circles do not have 5′-hydroxyls, the ability of these bands to be radioactively labeled indicates that they are not circles or lariats. Interestingly, a SYBR™ Gold stain reveals that RNase R digestion of the recombinant products leaves a band at about 27 nucleotides (Figure 47). This indicates that a lariat RNA species may be present amongst the gamma and/or gamma-prime bands which would not be seen in the radioactively labeled RNA. In any case, the RNase R digestions convincingly show that the gamma and gamma-prime bands are not circular RNAs.

In addition to RNase R digestions, we investigated the digestion of **R16** products by means of RNase A/T1. RNase A is an endonuclease that cleaves at YpN bonds while T1 is an endonuclease that cleaves GpN bonds – the combined mixture of these RNases should degrade the linear **R16** products to monomers and dimers. However, the branch junctions in branched RNAs are unreactive to RNase A/T1 treatment and should be left over as products of a complete digestion.

The action of RNase A/T1 can be mimicked with magnesium, basic pH, and heat, which accelerates internal cleavage. Accordingly, we reacted gel-purified **R16** recombinant products for three hours in a mixture of RNase A/T1 alongside a separate reaction in which we treated the recombinant products for three hours at 90°C with 50 mM $MgCl_2$ at pH 9.0. We then compared the results of each digestion to a complete

digestion of unreacted **R16** (Figure 48). In each case, we radiolabeled the mixture after digestion, which is necessary to visualize the internal nucleotides and branched remnants of complete digestion, and analyzed the reaction on a 22% polyacrylamide gel. In both cases, digestion of the recombinant products leaves traces of three distinct but closely packed bands that could be radiolabeled trinucleotide branch junctions, which cannot be cleaved into nucleotides by internal cleavage.

4.6 Adapter ligations and gel mobility shift

If the gamma and gamma-prime products are in fact branched RNAs, they should have multiple 3′ terminals. If so, the addition of adapters to the 3′ ends of a branched RNA should produce multiple gel shifts from a single ligation event. We therefore designed an experiment to ligate an adapter to the 3′ ends of possible branched RNA (Figure 49). We used a DNA adapter with a dideoxy terminal to prevent self-ligation and multiple additions, and we used the enzyme T4 RNA Ligase 2 (NEB) which only takes as a substrate an adenylated 5′ adapter, to eliminate the possibility of ligating product RNA. We ordered the adapter from IDT with a phosphorylated 5′ end, preadenylated the adapter according to a published procedure[40] and gel purified the slightly larger adenylated product on a 22% gel.

With our devised ligation procedure, we performed a ligation on the gel-purified recombinant products with excess adapter (Figure 50). The results provide powerful evidence for the existence of branched RNAs – there is a substantial shift upwards and increase in number of products after ligation that would not be expected from a single adapter addition to a linear product RNA. Multiple controls confirm that the enzyme and adapter have no reactivity by themselves – the dideoxy terminal cannot be ligated – and

the reaction with only enzyme and product RNA does not generate additional products since the recombinant products are not adenylated. Finally, the ligation with unreacted **R16** does not produce products larger than about 37 nucleotides, the size to be expected as a result of ligation of the adapter to R16. The overall result is replicated when the ligation products are phosphorylated with γ-^{32}P ATP (Figure 51) and visualized by phosphorimaging instead of SYBR GOLD™ staining.

The individual products of the **R16** reaction were gel-purified and the adapter ligation was performed on the purified RNA products to identify which **R16** reaction products were responsible for the products of the adapter ligation. The ligation products were 5′ radiolabeled and analyzed by PAGE, which shows that both the gamma (Figure 52) and gamma-prime products (Figure 53) have ligation products at or in excess of 150 nucleotides long, almost three times the size that would be expected if these RNAs were linear. Furthermore, the results indicate that the gamma and gamma-prime bands have multiple ligation products, an indication of branched RNA in which not all of the branches have been ligated to an adapter. Under the conditions of these experiments, neither circular RNA nor lariat RNA would be visible since they lack the necessary free 5′-OH needed for ^{32}P labeling. A SYBR™ Gold analysis of the ligation reaction of the gamma products reveals a similar product profile to that observed with ^{32}P labeling (Figure 54). In contrast, the ligation reaction products of alpha-prime and beta do not show products above the expected size for a linear RNA (data not shown), confirming that the gamma and gamma-prime bands are responsible for the gel shift and branched species.

4.7 Generalization of alpha-prime recombination

As a final test of the prebiotic utility of our recombination scheme, we attempted to generalize the reaction. Firstly, we identified the purpose and value of each nucleotide of R16 to create a general setup for alpha-prime self-recombination (Figure 55). We then wrote a brief computer algorithm to iterate through the entirety of 16mer sequence space and find all possible oligomers that matched this setup. This examination reveals 7,077,888 oligomers of size 16 that meet all the requirements for an alpha-prime structure, equating to about 1/606 or 0.16% of the total sequence space. Although some of the sequences that meet these criteria (such as AAAUUAAAAAUUAAAA) are not experimentally useful due to base composition, a majority of them are reasonable for randomly generated 16mers.

Based on our examination of 16mer sequence space for potential self-recombining alpha-prime setups, we chose one 16mer identified during the simulation (F16, 5′-CACGUAGUCCGGCUCA) which has an alpha-prime setup similar to **R16** (Figure 56) and designed a closely related second oligomer B16 (5′-CAGCUUAGUCCGGUUC), which had a four-nucleotide bulge between base-paired regions and a three-nucleotide tail (Figure 57). We purchased these oligomers, gel purified them, and reacted them under our standard recombination conditions. We found that with B16, there are no apparent alpha-prime or beta products in the reaction. However, B16 does generate products with mobilities similar to the gamma and/or gamma-prime products of the **R16** reaction (Figure 58). F16 generates a product profile whose bands are similar in size to the expected alpha-prime and beta products of **R16**, which migrate faster than the products of the B16 reaction. (Figure 59).

4.8 Conclusions

In summary, we have characterized both linear and non-linear recombination products of **R16** and **H13** self-recombinations and identified the 5′ cytosine hydroxyl as an important nucleophile for production of alpha-prime recombination products, as predicted by both our own[39] and other experiments with random RNA recombination[51]. The nucleophile almost certainly attacks a 2′-3′ cyclic phosphate ester formed by cleavage initiated by the adjacent 2′-OH, meaning that the general process for RNA recombination with 5′ nucleophiles requires cleavage as an activating step. The cleavage is also specific to its position in the secondary structure of bound RNAs; in the alpha-prime process, maximal cleavage over the 3-nucleotide bulge occurs so that only one or two nucleotides are linked over the bulge after the 5′-OH attack and ligation event. Thus, for the major linear product of **H13**, we observe the loss of the terminal G followed by attachment of a 5′ cytosine to create a one-over-three internal loop in the splint-substrate double strand, and for **R16**, the most cleavage happens at the 12-U position resulting in a two-over-three internal loop after ligation. Both phosphorylation of the 5′-OH and deoxyribose substitutions in the **R16** tail completely eliminate the alpha-prime linear bands, making a compelling case that this reaction is a two-step mechanism, with cleavage and ligation specific to certain junctions in the three nucleotide bulge, or in the case of the alpha reaction, specific to the displaced tail.

For the gamma and gamma-prime bands, we have gathered considerable evidence in support of a branch, in particular from the adapter ligations that produce and upward distribution of products, and the radioactive labeling of 5′ hydroxyls which would not be possible with lariats or circles. We have also demonstrated that the reaction for these

products is rapid, with no intermediates, ruling out a multi-step recombination process and implying a one-step recombination, unlike our linear products, that may take place with a 2′ internal hydroxyl nucleophile. In addition, RNase R digestion reduces these bands to bands in the linear region (27-29 nt), which would not be possible with a circular species.

Identifying a nucleophile for the reaction that leads to the gamma and gamma-prime bands has proved to be difficult, and the revelation that a DNA construct of **R16** produces a product band in the gamma region suggests that a non-canonical branch or linkage may be possible. Elucidating the nature of such a linkage falls outside the scope of our investigation, however there are still some conclusions we can draw. The presence of the gamma and gamma-prime bands in the R16-d{AUUU} construct is inconsistent with a nucleophile in that region being necessary for the gamma and gamma-prime bands, eliminating the possibility of a Silverman-type reaction[41, 42]. From the results of our high-throughput sequencing, it appears that many of the anomalies (insertions, deletions, mutations) in sequenced **R16** recombination products are present at two different junctures, one in the splint between 5C and 9U, and another at 10G or 11C. At the former of these junctions, a C insertion is frequently observed prior to 7G, whereas at the latter, either a C, or predominantly, a G insertion is observed on or after 10G.

From our results in Figure 39 that show approximately 5 gamma bands in the **R16** reaction, we can surmise that the close packing of these bands is the result of a similar piece of RNA being attached to a different juncture. On the basis of our sequencing data and from the results of recombination of other deoxy residues, we propose that up to five 2′-hydroxyl nucleophiles in the splint are attacking the junction between 10C and 11U to

add five or six-piece branches to **R16** in the splint region between 5C and 9U. Our evidence for this junction comes from the one-hour radioactive reaction, in which a strong radiolabeled band at 10 nucleotides is consistently seen in conjunction with the appearance of recombinant products, and from digestions of product RNA with RNase A/T1 and alkaline hydrolysis. Both of these latter treatments, in which RNA was radiolabeled after digestion, show 6mer bands that would be the result of fragmentation of such a branch, and the lack of bands above or below them indicates that these bands must be generated continuously during digestion, as would be expected if transesterification products were hydrolyzed in the transition state.

We tested the possibility of splint nucleophiles by investigating the reaction of modified **R16** with deoxyribose nucleotides at positions 7G, 8U, and one with three deoxyribose nucleotides at 6C, 7G, and 8U. The constructs with d7G or d8U do not produce any observable change in the self-recombination reaction. The use of a triple-deoxy construct with 6C, 7G, and 8U all containing deoxyribose nucleotides does produce an observable change in the gamma region, but does not affect the reaction that forms the gamma-prime product. An explanation for this result is that multiple nucleophiles are responsible for the products in the gamma region and those nucleophiles can attack at multiple junctions such that the loss of one nucleophile does not produce any apparent change in the gamma region. In contrast, the loss of multiple nucleophiles does produce an observable change in the gamma region and our results show that the loss of three 2′ hydroxyls in the bulge region of the splint does appear to cause the loss of some products in the gamma region.

It must be noted that the deoxy splint causes a shift of the linear products – the major product is now 28 nucleotides instead of 29 – an illustration of how deoxy substitutions can affect secondary structures. It is not likely that this is a mobility shift because the gamma-prime product of this construct migrates at a nearly identical rate to the gamma-prime product of the original **R16**. Thus, while we conclude that at least some nucleophiles are in the bulge region of the splint, more convincing analysis by mass spectrometry or NMR is likely necessary for full confirmation.

The difficulty of analyzing the gamma and gamma-prime products is underscored by the fact that a DNA construct of **R16** also produces a band that is in the gamma-prime region. As the RNase R digestions convincingly rule out circular RNA and the product is more than twice the size of the starting material, it is unlikely that a reaction involving a sugar hydroxyl produces this product; *cis*-attack of a terminal 3′-OH on an existing phosphor-ester linkage with a nucleoside or polynucleotide leaving group could only produce a circle. In short, it seems possible that this is not trans-esterification chemistry; our transcribed versions of **R16** all manifest this band so it is unlikely to be contamination from the synthetic process used to make the RNA. Rather, it seems that this is a type of non-canonical RNA-catalyzed chemistry, such as a linkage that might be formed from a nitrogenous nucleophile in a nucleobase.

The final question to be addressed is whether the R16-d10G and R16-d11C oligomers produce an observable change in the self-reaction, as suggested by our high-throughput sequencing. The R16-d10G reaction results are not significantly different than the original **R16**, and all of the gamma and gamma-prime products can be observed. While the gamma-prime and gamma products are similarly unchanged, the d11C reaction

appears to be missing the linear beta recombination product at 31 nucleotides, an unusual result considering that only an internal 2′ hydroxyl has been changed.

Since the RNase R digestions of the gamma and gamma-prime products eliminate all but one of these two bands, it is possible that there is at least one lariat RNA species present in the gamma region. In addition, we note in the HTS sequencing data that there is a large amount of R16 lacking the 5′ cytosine, an indication of specific cleavage at this juncture. Therefore, we propose a two-step mechanism for forming the beta products of **R16** (Figure 60) in which the 2′-OH of 11C attacks the phosphodiester bond between 1C and 2G to form a small lariat. Re-attack of this 2′-5′ lariat linkage can occur when another molecule of **R16** binds to the tail of the lariat. The terminal guanosine is hydrogen bonded to the same cytosine on which the branch occurs, and its 3′-OH may be able to attack the branched juncture to resolve to a linear product. This theory is consistent with the RNase R result and gives a possible unique mechanism for the beta result.

Finally, our discovery of novel recombination mechanisms for small RNAs is prebiotically significant because it demonstrates how length expansion can occur among very small RNAs, possibly leading to the formation of ribozymes and novel RNA structures out of random precursors. The alpha-prime mechanism shows that energetically neutral template-mediated cleavage and ligation reactions can form longer RNAs that may fold into more structured molecules. In some cases, these longer molecules may be able to react multiple times, producing even longer RNAs. While it is less clear how branched RNA and lariats (if formed) would be prebiotically useful, such molecules could still function as templates or other ribozymes, and their existence demonstrates that internal 2′ hydroxyls can participate in transesterification reactions of

even small RNAs. Finally, the beta mechanism is useful as a potential means of generating recombinant oligomers without the 5′-OH nucleophile that is necessary for the two-step cleavage and ligation of the alpha-prime process. This mechanism is perhaps the weakest mechanism in terms of yield, but it nonetheless demonstrates that the terminal 3′ and 2′ hydroxyls can also engage in transesterification reactions among short RNAs.

Chapter 4 Figures

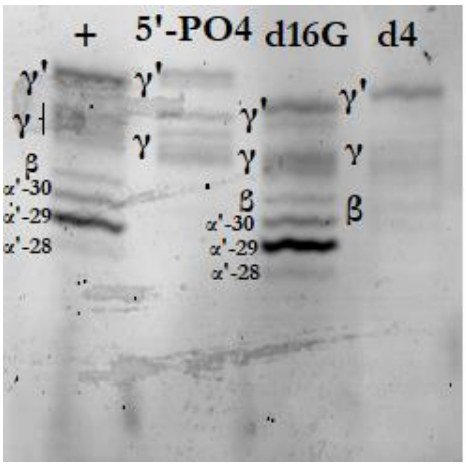

Figure 39: Reactions with modified R16 oligomers. {+} indicates original R16, followed by 5′ phosphorylation, a terminal deoxyG, and d4, which represents d{AUUU) or a version of R16 with deoxy residues at all four of the nucleotides prior to the terminal G.

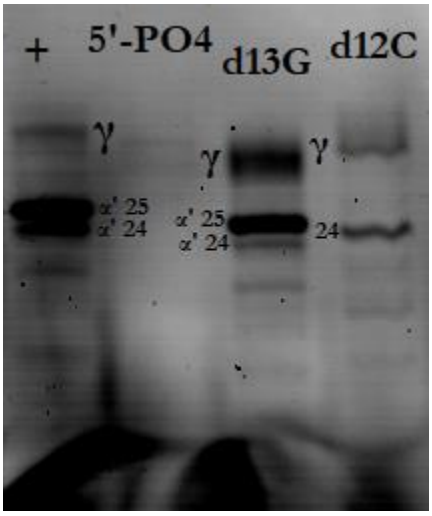

Figure 40: Modifications of H13. {+} represents unaltered H13, 5′-PO4 indicates 5′ phosphorylation, and d13G and d12C refer to oligomers with deoxy residues at those positions. In case of the former, this is the terminal 2′-OH which is absent; in the latter, the penultimate cytosine contains deoxyribose.

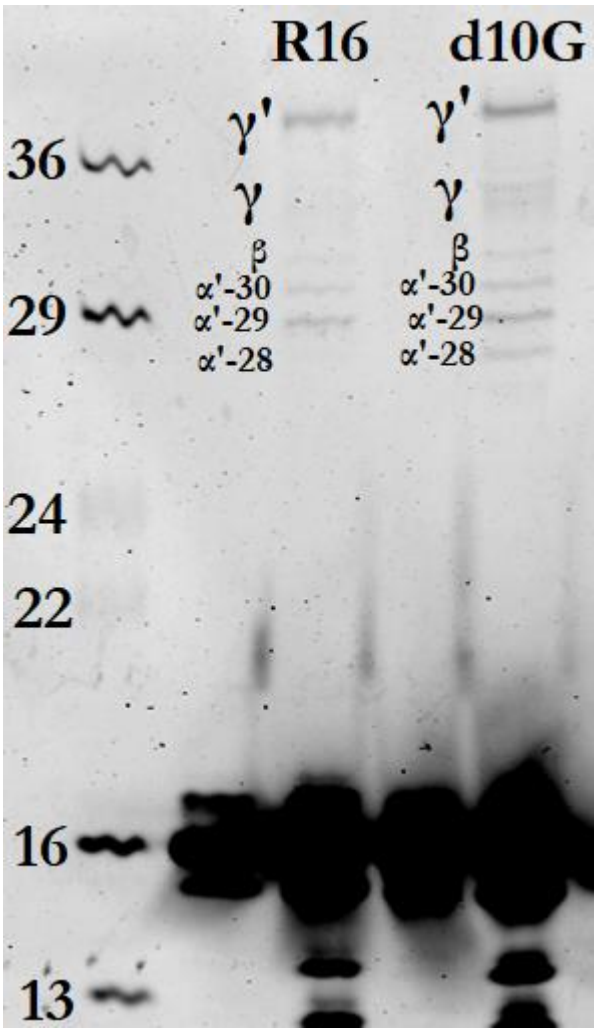

Figure 41: Modification of R16 with a deoxyG at position 10 does not affect the reaction.

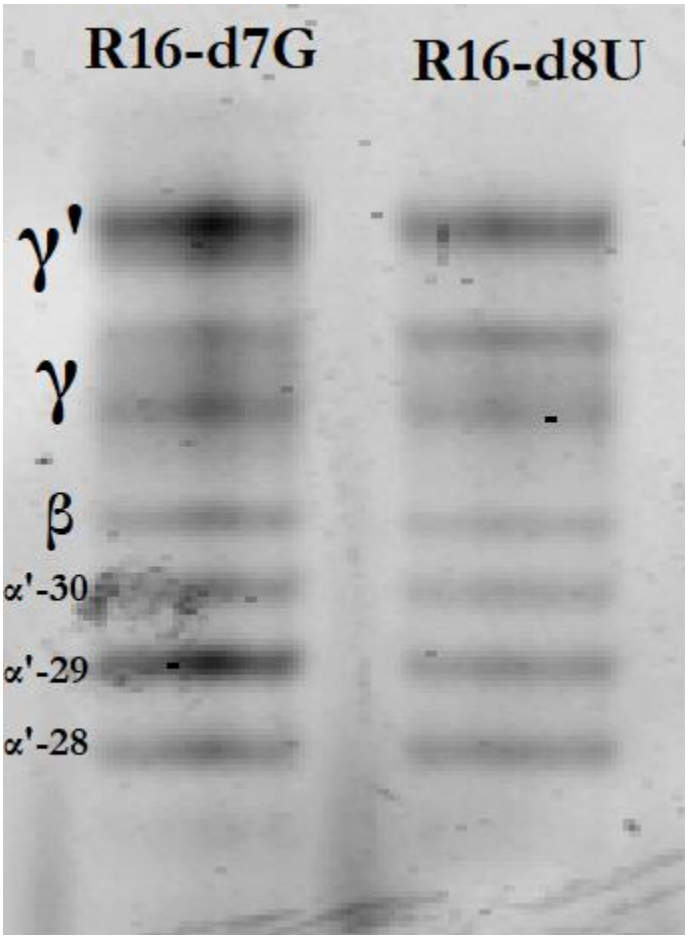

Figure 42: R16 constructs with deoxy residues at 7G and 8U, which were in the splint region implicated by HTS evidence as possible branch sites.

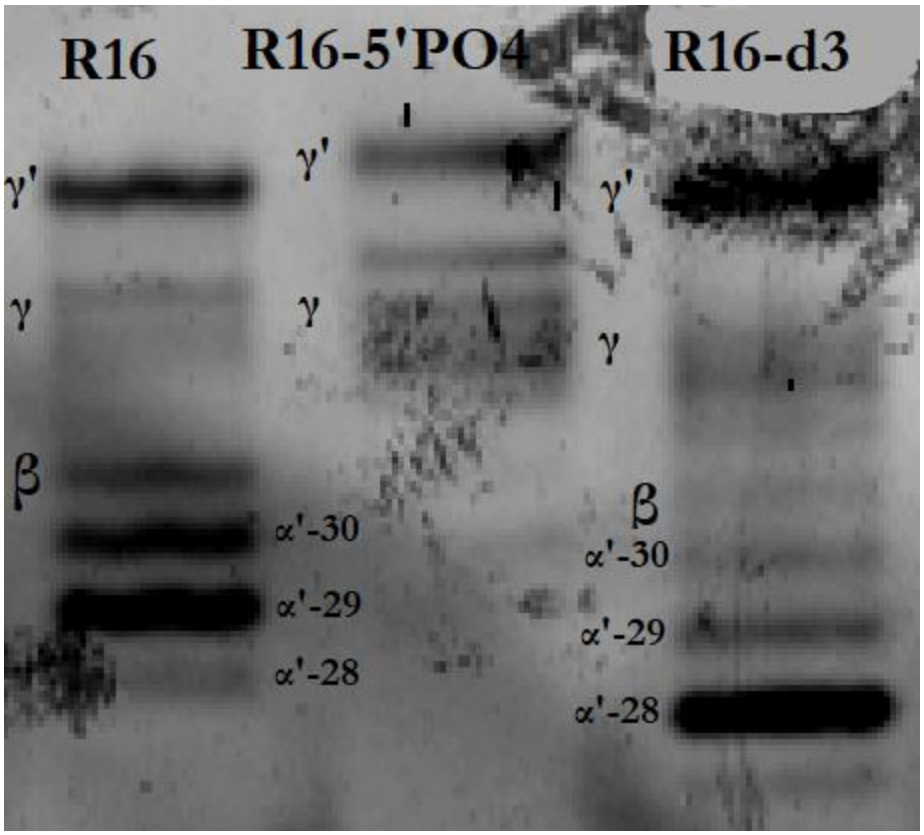

Figure 43: Triple-deoxy construct of R16, with three deoxy residues in the bulged splint region of the alpha prime triplex.

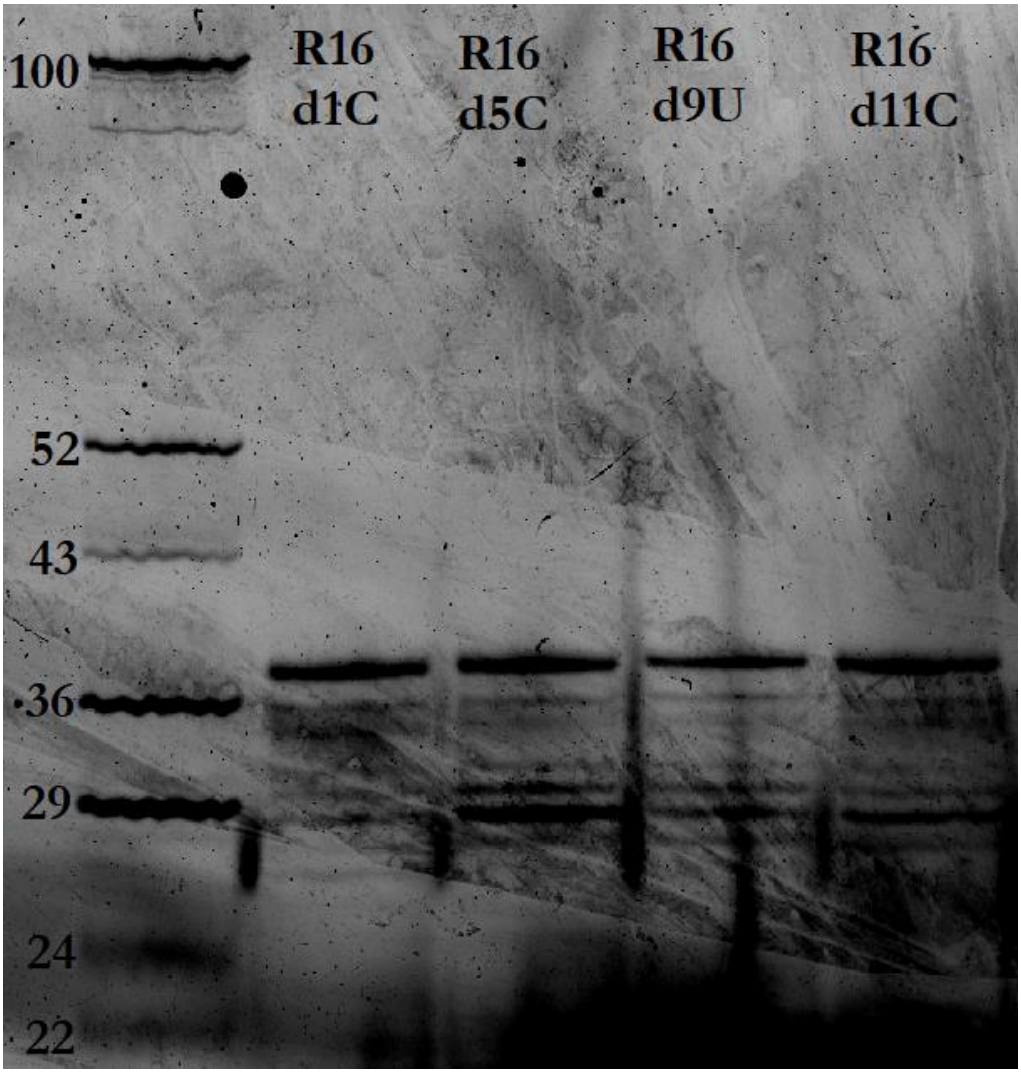

Figure 44: Self-recombination of R16 deoxy constructs with deoxyribose at 1C, 5C, 9U, and 11C.

Figure 45: An all-DNA version of R16 produces a faint band consistent with the gamma-prime band in R16. Left: High contrast version showing products. Right: Low-contrast version showing starting material. The DNA migrates 2-3 nucleotides faster than the RNA, so while this band parallels the gamma bands, we think it more likely to represent the gamma-prime band.

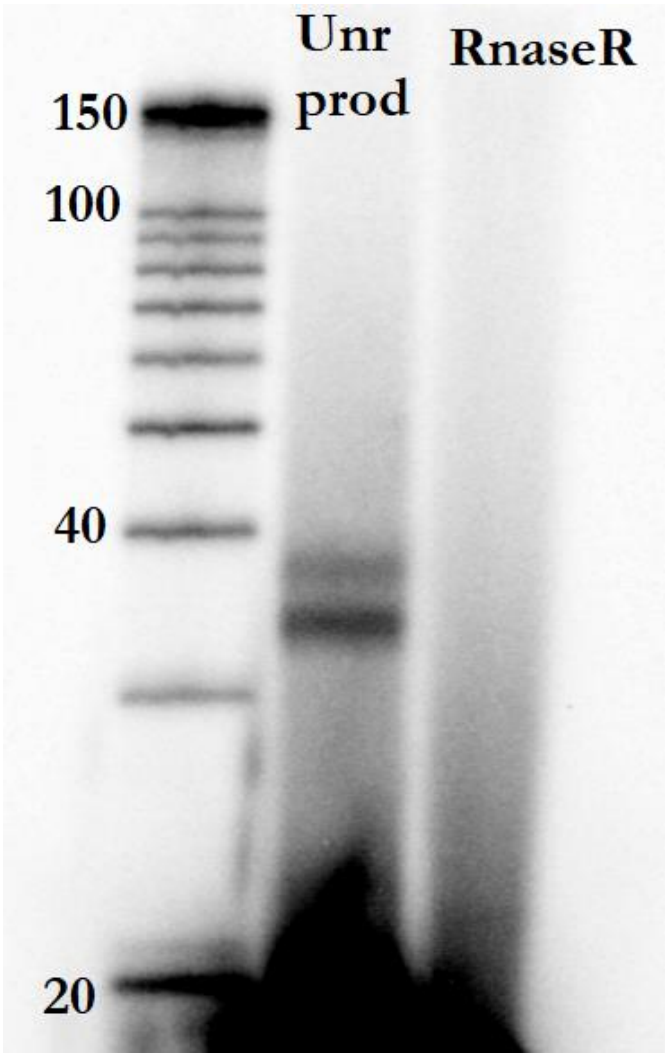

Figure 46: RNase R digestion of radioactively labeled gamma and/or gamma-prime products. A radiolabeled ladder is shown on the far left, while the middle is the positive control including the radiolabeled gamma and/or gamma-prime bands.

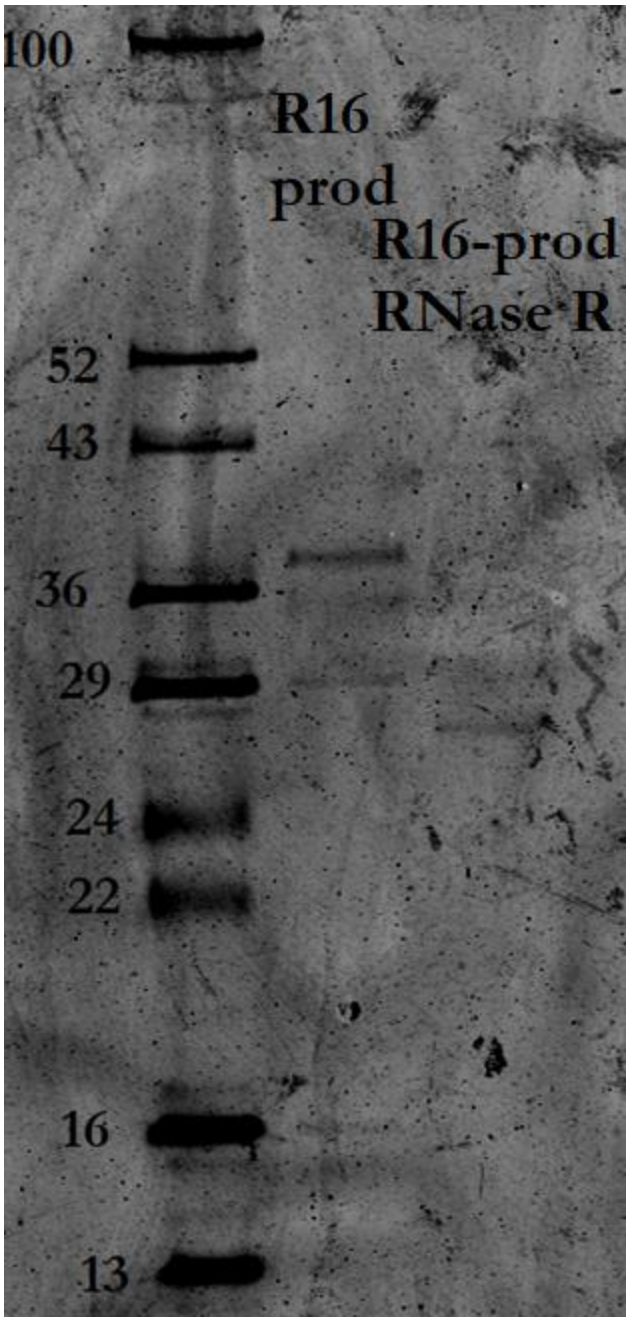

Figure 47: RNase R digestion of seven-day reaction products, with visualization by SYBR™ Gold.

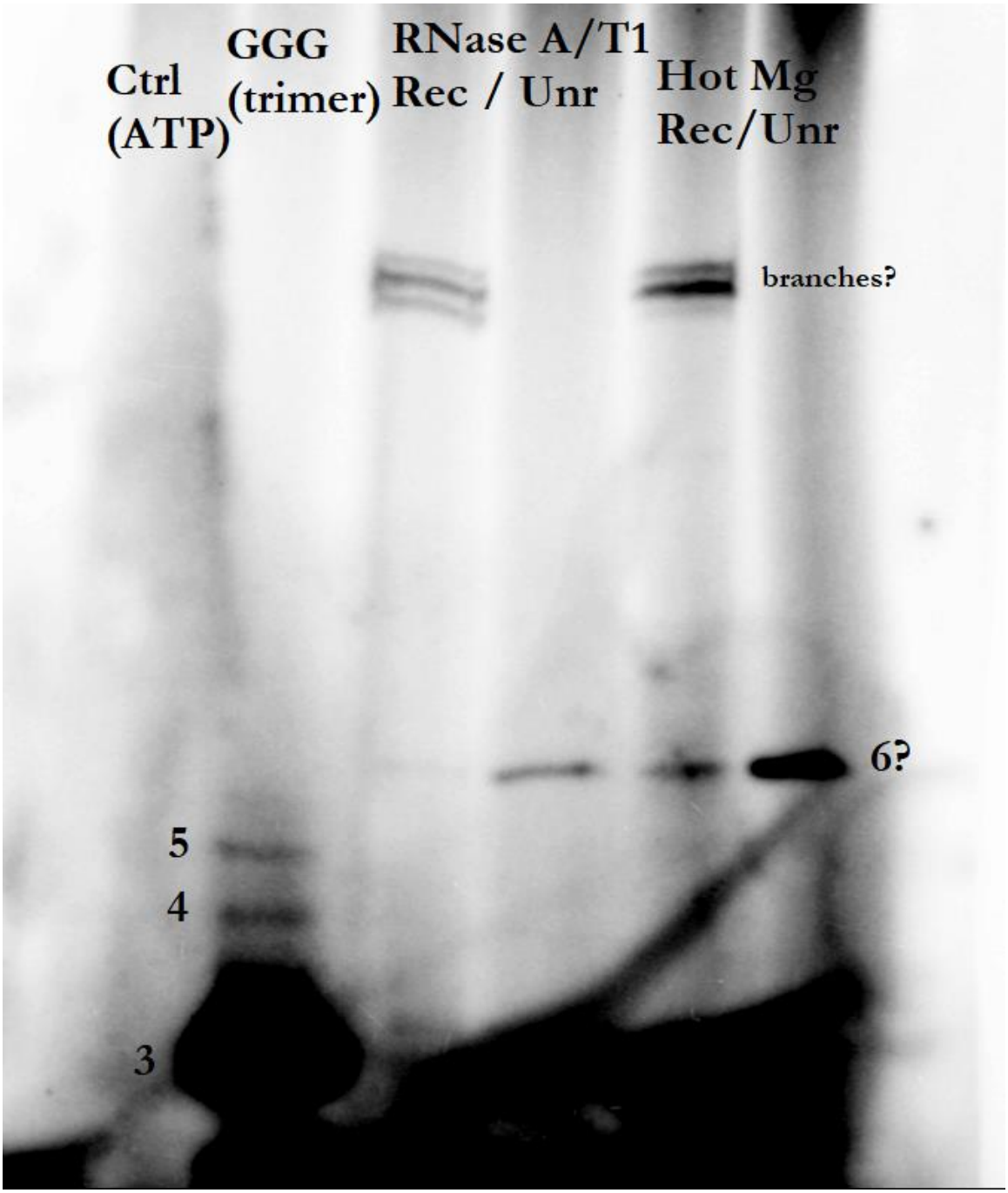

Figure 48: RNase A/T1 and hot basic digestions of unreacted R16 and R16 products. “Rec” represents digestion of recombination products while “Unr” represents digestion of unreacted R16.

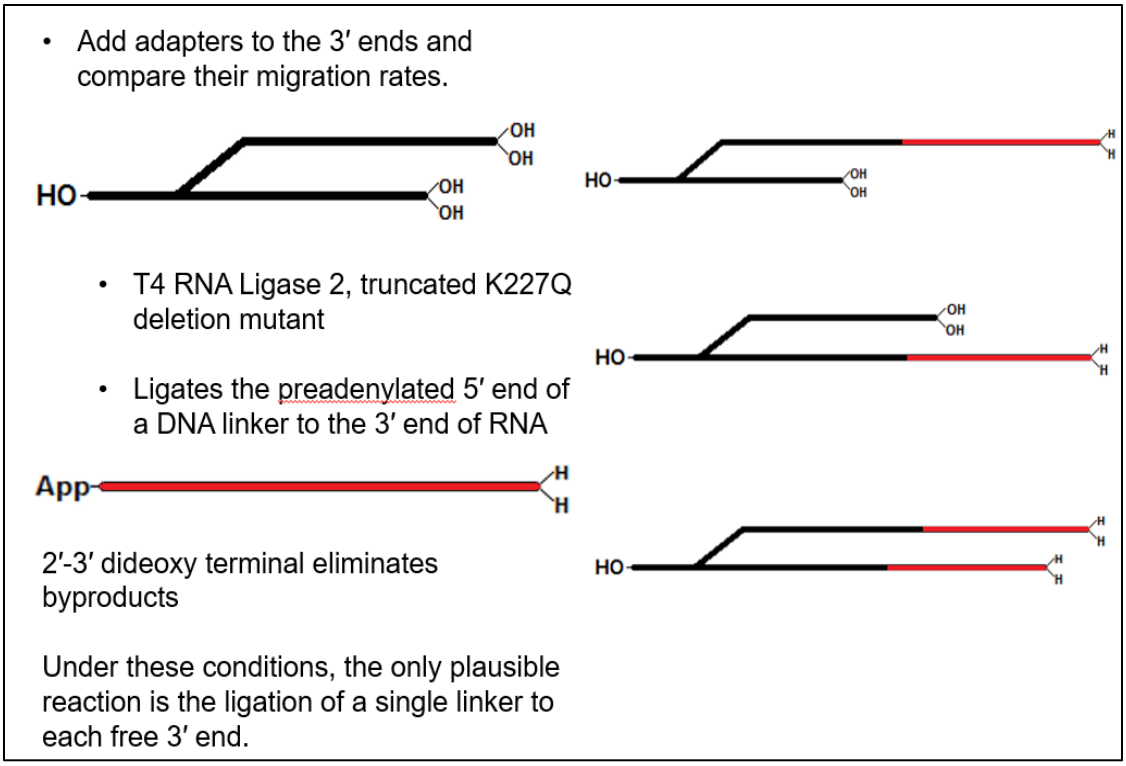

Figure 49: Scheme for ligating a 21-nt DNA adapter to branched RNAs.

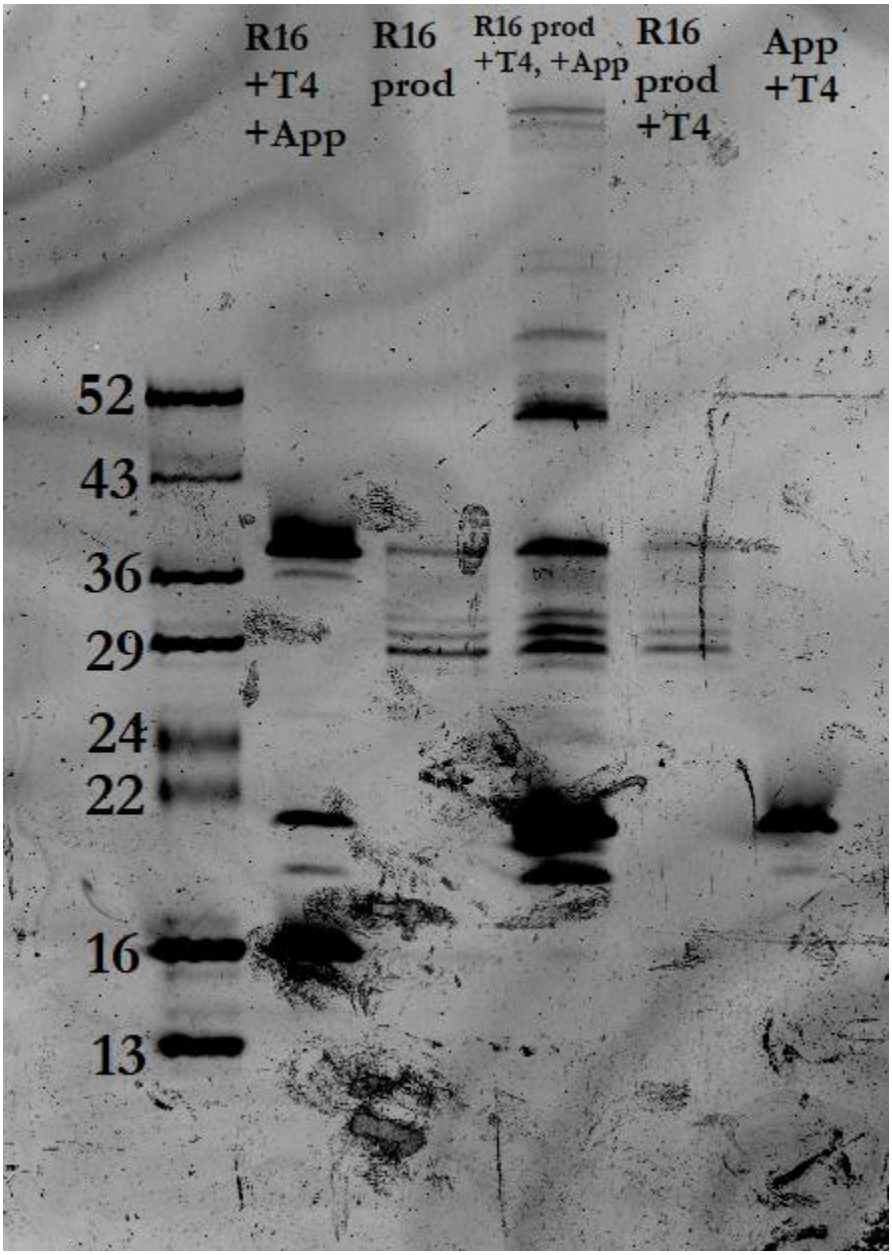

Figure 50: Adapter ligation of R16 recombination products. Left to right: (1) ladder, (2), unreacted R16 with adapter and enzyme treatment, (3) R16 products without treatment, (4) R16 products with adapter and enzyme, (5) R16 products with only the enzyme, and (6) the adapter with only the enzyme.

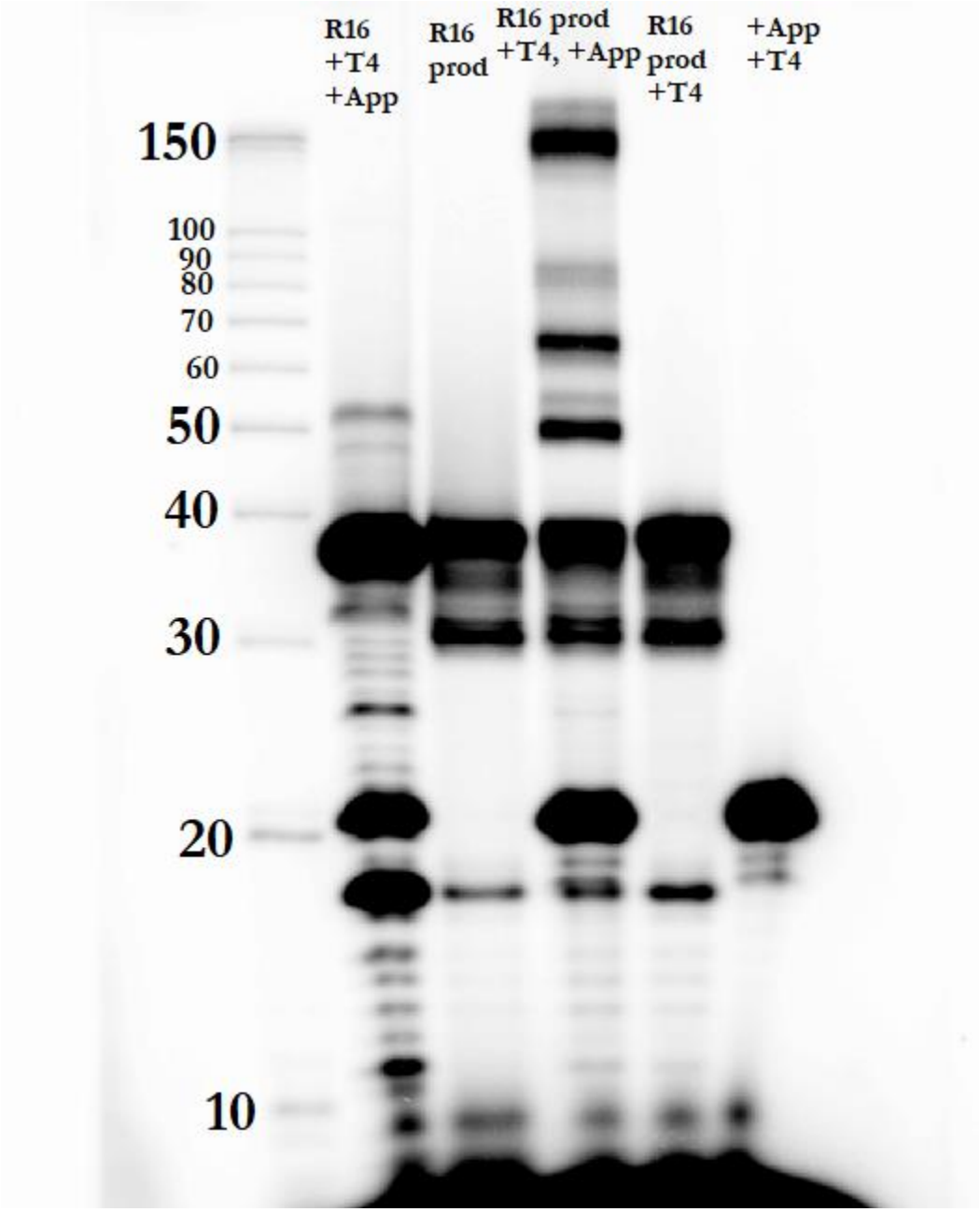

Figure 51: Radiolabeled adapter ligation. The designation “Prod” refers to the gel purified R16 recombination products. Left to right: (1) ladder, (2), unreacted R16 with adapter and enzyme treatment, (3) R16 products without treatment, (4) R16 products with adapter and enzyme, (5) R16 products with only the enzyme, and (6) the adapter with only the enzyme.

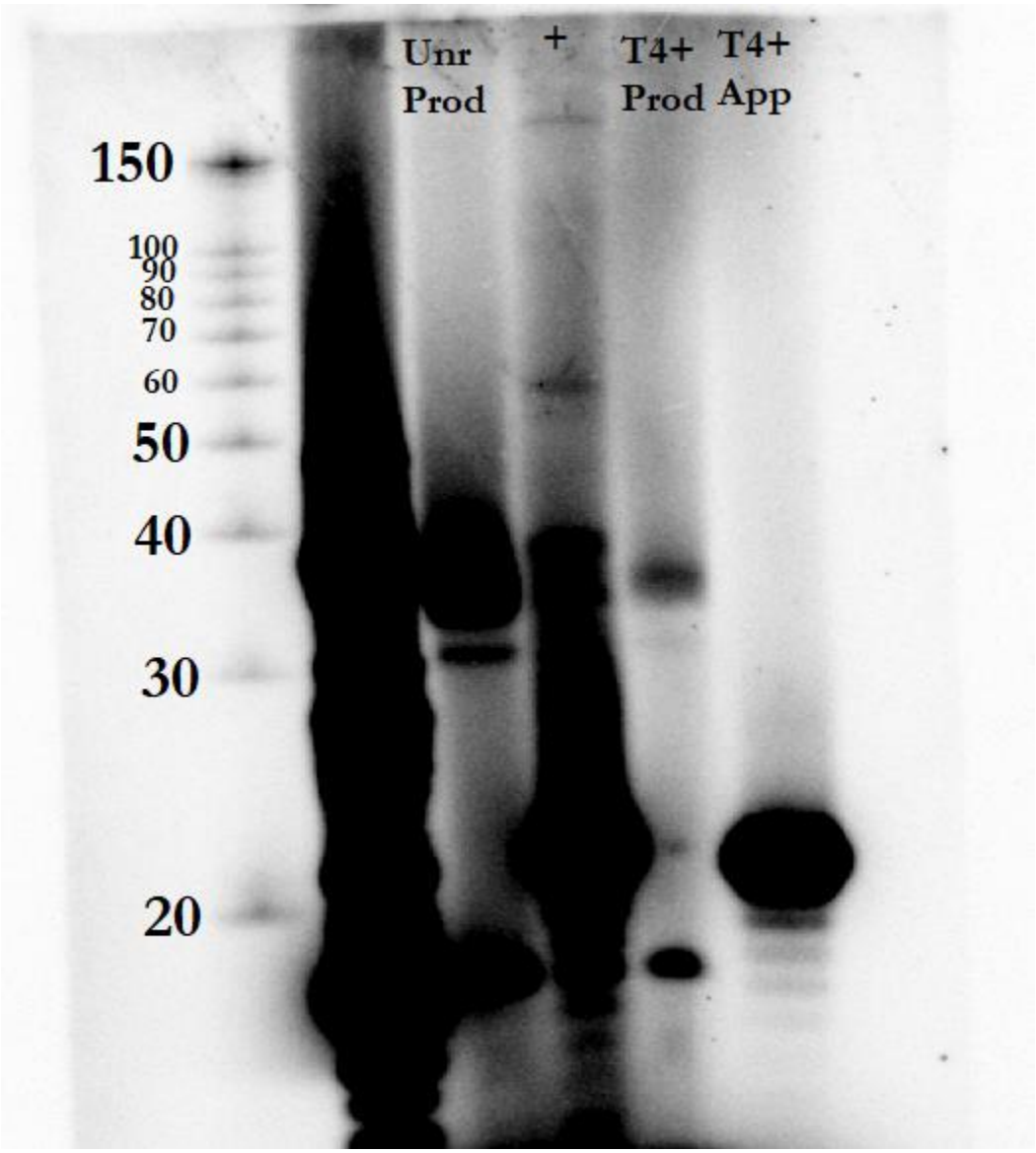

Figure 52: Radiolabeled adapter ligation exclusively of gamma bands. The {+} lane is the products with the addition of adapter and enzyme. Note that the ladder in the leftmost lane is terminated with a cyclic phosphate and thus migrates slightly faster than the RNA. Left to right: (1) ladder, (2), unreacted R16 with adapter and enzyme treatment, (3) gamma products without treatment, (4) gamma products with adapter and enzyme, (5) gamma products with only the enzyme, and (6) the adapter with only the enzyme.

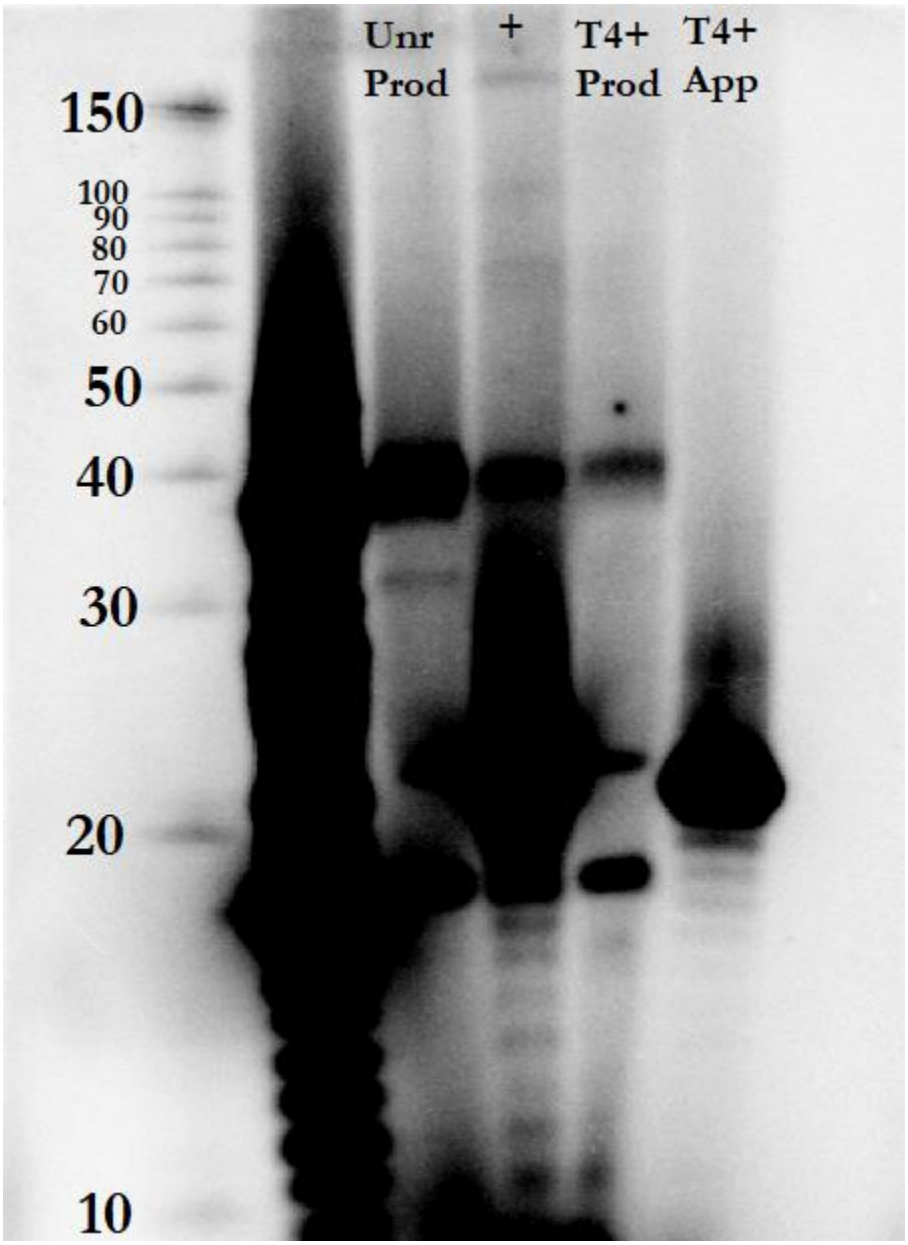

Figure 53: Radiolabeled adapter ligation for gamma-prime bands. The {+} lane indicates product plus enzyme and adapter. Left to right: (1) ladder, (2), unreacted R16 with adapter and enzyme treatment, (3) gamma-prime products without treatment, (4) gamma-prime products with adapter and enzyme, (5) gamma-prime products with only the enzyme, and (6) the adapter with only the enzyme.

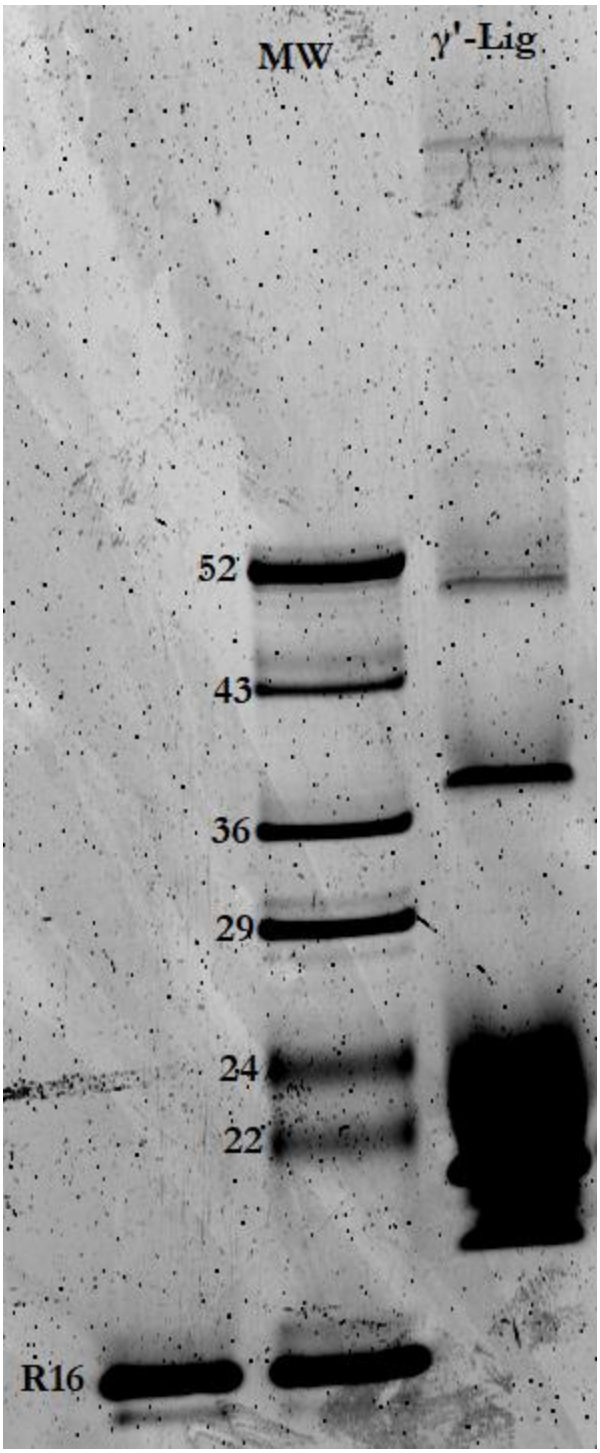

Figure 54: SYBR™ Gold stain of gamma-prime adapter ligation. Left to right: (1) R16, (2) ladder, (3) R16 gamma-prime recombination products with adapter and enzyme.

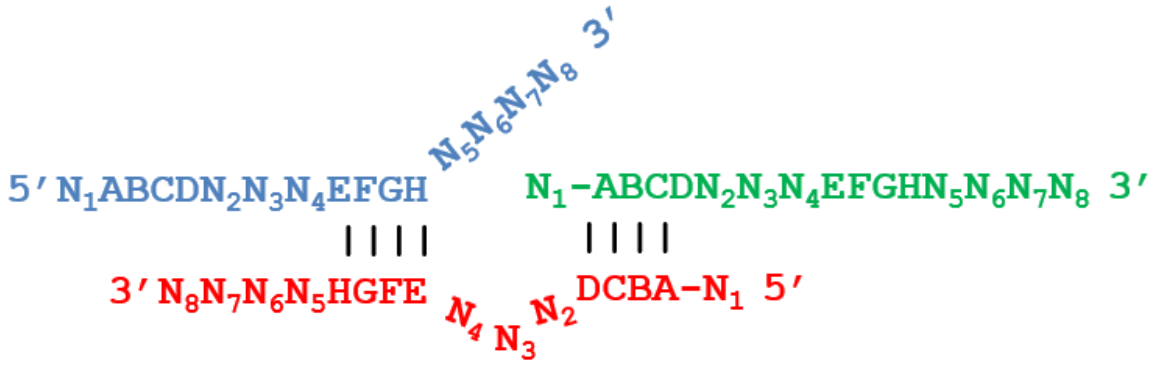

Figure 55: Generalized scheme for alpha-prime recombination of 16mers. A is complementary to D, B is complementary to C, E is complementary to H, and F is complementary to G. In addition, N5 cannot be complementary to N4, N3 cannot be complementary to N6, and N2 cannot be complementary to either N1 or N7.

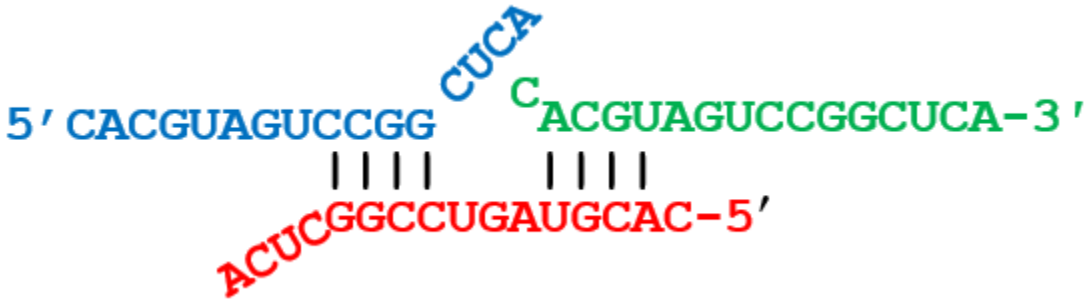

Figure 56: Alpha-prime triplex of F16

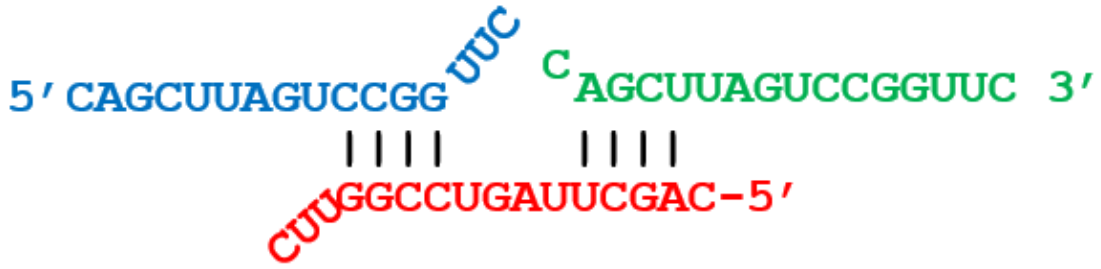

Figure 57: Theorized alpha-prime triplex of B16 with extended bulge and shortened tail.

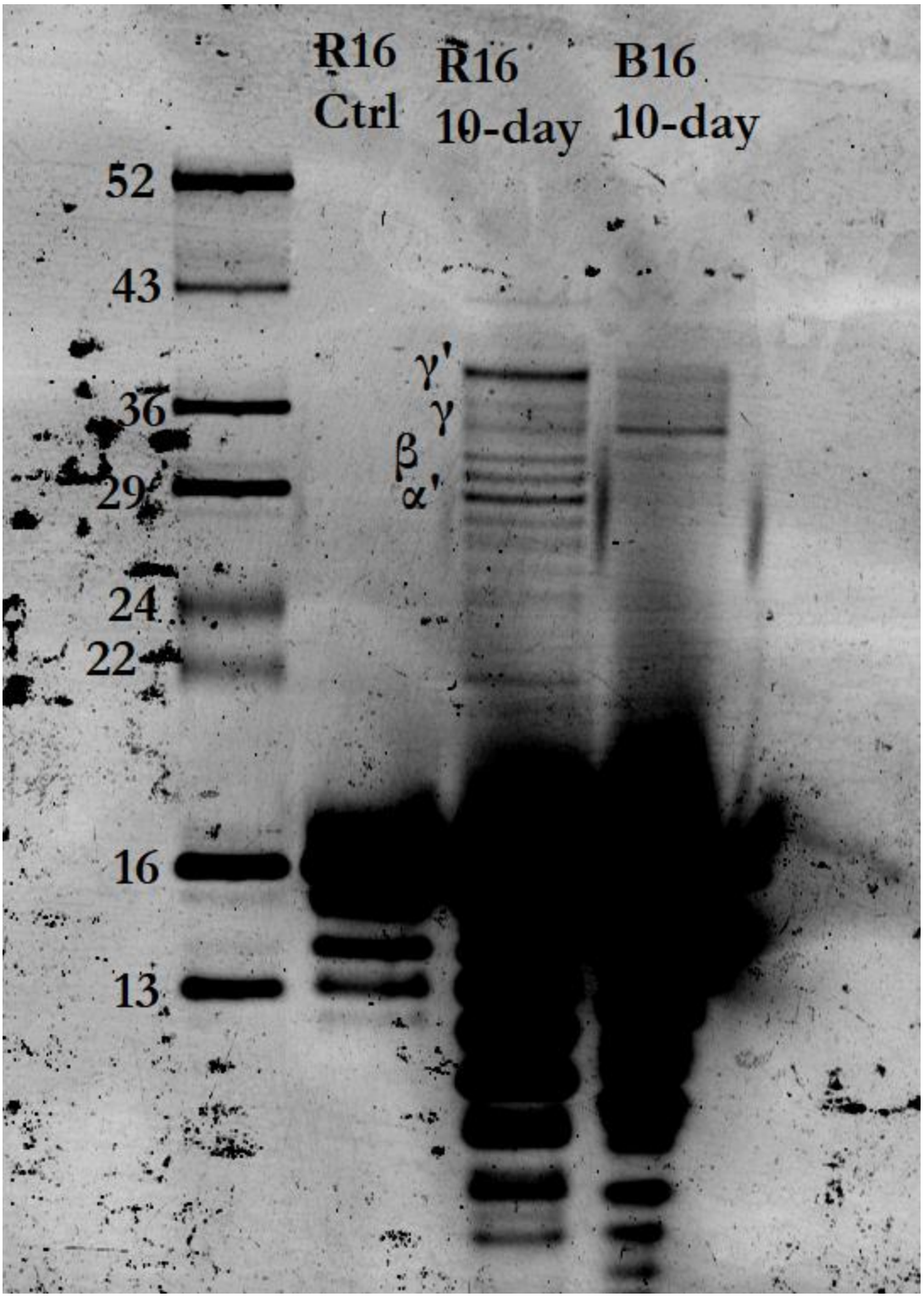

Figure 58: B16 as compared to R16. Both RNAs were reacted for 10 days of cold cycling at pH 8.0 with 100 mM magnesium.

Figure 59: Comparison of B16 and F16. Left to right: (1) Unincubated B16 as a control, (2) B16 reacted for seven days under standard conditions, (3) unincubated F16, (4) F16 incubated under standard conditions for seven days.

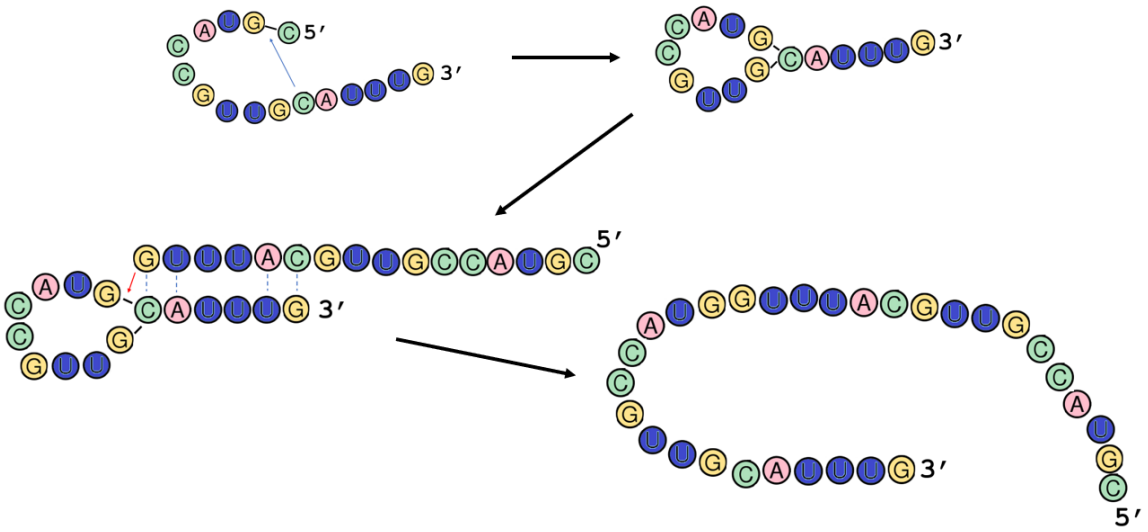

Figure 60: Proposed two-step mechanism for the linear 31mer recombination product “beta.” The blue arrow indicates attack of an internal 2′-OH onto a phosphodiester bond while the red arrow indicates attack of a terminal 3′-OH onto the 2′-5′ linkage of the lariat. Dashed lines represent hydrogen bonds. A lariat is formed between 11C and the first phosphodiester bond, with the 5′-C as the sole leaving group. A second molecule of R16 binds to the 3′ tail of the lariat and is hydrogen bonded to the same cytosine on which the branch occurs – its terminal 3′ hydroxyl attacks the 2′-5′ branch to resolve to a linear molecule.

Chapter 5: Computer simulations of recombination reactions

In order to further explore the principle conclusion that recombination is an effective means of building up size, sequence diversity, and structure in a putative prebiotic mixture of small RNAs, we constructed sequence-explicit computer simulations of alpha recombination, the joining of strands initiated by cleavage followed by attack of 5′-hydroxyls on 2′-3′ cyclic phosphates, which is likely the most predominant form of recombination in random or semi-random pools.

There were two primary goals of computer simulations for this work. It should confirm the experimental observation that spontaneous, random cleavage and ligation reactions can alter the distribution of a short oligomer pool by biasing it towards both shorter and longer oligomers. Second, by adding feedback mechanisms, we investigated whether an increase in background cleavage and ligation rates caused by transient ribozymes could accelerate recombination enough to form oligomers in the range of 100-200 nucleotides long, the length required to form complex catalysts.

Although there have been many attempts to create simulations of life or life-like processes, especially in the field of evolutionary biology, modeling the dawn of the RNA World has been a relatively recent and somewhat obscure endeavor in computational biochemistry. Hogeweg has used cellular automata to model the emergence of replicating structures[52] and Dyson proposed a toy model in which ordered, mutually catalytic molecules emerged from disordered initial conditions[53]. In 2002, Szabo et al. created a model showing how polymerase replicators with specific sequences could emerge and persist with better fidelity, templates, and length[54], but their abstract model assumed the

ability of small ribozymes to catalyze replication and skirted around the realities of known chemistry.

Following the previous models, Wu and Higgs developed a model demonstrating that the spontaneous emergence of an RNA polymerase ribozyme coupled with autocatalytic feedback could act on a pool of RNA oligomers to generate a "living" state containing robust catalytic polymers and a high level of activity, as opposed to a "dead" state in which there are few ribozymes and little to no activity[55]. In this model, a finite pool of RNAs is represented by differential equations corresponding to their characteristics and activities. There is no sequence information and activity of the oligomers is inferred by length. Additional models by Higgs have examined phenomena such as homochirality[56], diffusion and wet-dry cycling[57], and compositional inheritance[58] with a basis in known chemistry. All of these models make the principle assumption that a polymerase or replicase ribozyme is essential to jumpstart the RNA World.

Wu and Higgs have also adapted their original model to examine the effects of nucleotide synthesis and reversible recombination in their polymerase-based model[59]. One key conclusion of theirs is that recombination will redistribute fixed-length oligomers into an exponential distribution with the same average length, but that includes some greater lengths than the starting material, with the majority of lengths being smaller. This model is partially based on the alpha recombination model of Lutay et al.[38], except that it is treated as fully reversible. As with prior models, there is no sequence information nor consideration of structure, sequence or stability of the chosen molecules. The authors conclude on this point that any closed exponential distribution is stable under recombination, although they ignore the phenomenon of spontaneous background

cleavage, which, in the absence of re-activation and re-ligation, will progressively reduce a solution of RNAs into monomers over long time periods. Incorporation of reversible recombination into Wu and Higgs's polymerase model thus does not produce any significant change – the authors determine that if recombination is fully reversible, there can be no autocatalytic state initiated by recombinases.

Wu and Higgs also consider whether recombination may, in some circumstances, be non-reversible, using as an example the autocatalytic self-assembly of the *Azoarcus* ribozyme demonstrated by Hayden and Lehman[34]. In this case, the formation of the *Azoarcus* complex from its substrate precursors is not fully reversible because the complex itself has a higher stability. However, because the *Azoarcus* ribozyme can only assemble its own substrates and cannot create them, improvement of the ribozyme only results in more efficient assembly of its substrates. The ability of the ribozyme to increase recombination of other substrates has no meaningful affect in this model because the authors previously determined that increasing the rate of recombination does not alter the exponential distribution.

Other models of polymerase-based RNA World emergence have been constructed by Ma and Zhang, who simulated early RNA evolution on a mineral surface[60-61]. Their model is interesting for its use of spatial diffusion on a mineral grid, but assumes activity from very small replicase ribozymes, a similar problem to earlier simulations. A further model describes how nucleotide synthetase ribozymes may have emerged first in the RNA World[62].

Most if not all of the well-regarded RNA World simulations make use of a high-fidelity RNA polymerase ribozyme that is needed to preserve information. Szabo, Higgs,

and Ma all assumed that a non-infinitesimally small random proportion of relatively small RNA molecules (5-60 nt) could catalyze template-directed replication but there is currently no experimental evidence for this. Furthermore, although there are a variety of potentially prebiotic syntheses for short RNAs, there is no evidence that abiotic polymerization and ligation of nucleotides can produce long RNA molecules. Ferris and Orgel showed that polymerization of activated nucleotides on montmorillonite clays generates RNA oligomers up to 55 nucleotides in length[23] but these nucleotides were activated with imidazole, which may not have been a prebiotically relevant molecule, although some recent research has suggested that it is[63]. In any case, the bulk of these oligomers fall within the range of 20-30 nucleotides, and most likely do not meet the length, sequence, or structural requirements for ribozyme-catalyzed RNA polymerization.

Perhaps the most specific treatment of recombination in a model to date was done by Myszor and Cyran[64], who used the model of Ma et al.[60] to simulate recombination on a mineral surface. The authors do not consider either true sequence or structure of the oligomers, and the simulated recombination reactions are only loose analogs of known chemistry, including both the reactions of Lutay et al.[38] and Pino et al.[37] Furthermore, while the distribution used by the authors resembles a distribution of abiotic RNA polymerization, it is small, and not truly exponential. This model is unique for its suggestion that recombination can improve the lengths of RNA oligomers in solution, a different conclusion than the one reached by Wu and Higgs. In the Myszor and Cyran report, recombination of a small Poisson distribution can produce recombinant oligomers up to 200 nucleotides long, a majority of which are larger than the starting distribution. However, it is unclear if the authors considered the chemical realities of the

recombination reaction. For example, the two-step alpha reaction features a ligation step preceded by a cleavage reaction with a higher overall rate, which should cause a majority of recombination events to produce shorter strands.

5.1 New simulation of RNA recombination

Here we produce for the first time a full informatic simulation of random RNA strands undergoing spontaneous *alpha* recombination, a term we developed to represent the reaction of Lutay et al.[38], in which two 16-nucleotide oligomers recombine over a splint molecule to produce a 28-nt length product and a 4-nt leaving group. I will show here that given an exponential starting distribution of RNA oligomers ranging from 1-16 nucleotides in length, successive and random recombination events can shift the distribution to generate longer molecules approaching the order of magnitude necessary to be complex catalysts. The model is unique among previous RNA World simulations in that it imparts full sequence information for every RNA and nucleotide, and includes polarity, binding requirements, and simple structural analysis and classification of products. I also specifically examine the redistribution of activated and unactivated oligomers, and confirm that preactivation is an accelerant for recombination and length expansion of short RNAs.

By imparting structural classification to the products, we can distinguish between different types of secondary structures and assign them certain stabilities and/or activities. We allow small recombinase ribozymes that arise in solution to have small catalytic effects – they may increase the rate of specific cleavage or ligation of other RNAs undergoing alpha recombination. Our results show that structural feedback, in conjunction with an increased recombination rate, can further expand the lengths in a

solution of random RNAs and create a significant amount of structure, setting the stage for the emergence of more advanced catalysts.

5.2 Simulation methods

A nucleotide-explicit computer simulation was written in C++ and simulations were carried out either on an AVA Direct Dual 2.4 GHz processor desktop computer with 128 GB RAM or a refurbished HP 3.00 GHz desktop computer with 16 GB RAM. The simulation emulates the long-term effects of persistent alpha-recombination on a solution of random RNA molecules. RNA oligomers were designed as "RNAStrand" objects containing several properties: sequence (string), length (integer), cyclic phosphate (true for a 2′-3′ cyclic phosphate, false for no phosphate or non-cyclic phosphate) ribozyme activity (true or false), structure (enum: GNRA, UNCG, CUUG, AANA, Tetraloop [nonspecific], Pentaloop, Hexaloop, Heptaloop), and kcal (sum of hydrogen bonds of self-folded structure using the estimate of one kcal ≈ one H-bond; a C-G pair is 3, an A-U pair is 2, and a G-U pair is 1)

An algorithm was created to find any and all alpha-recombination setups from three distinct RNA strands. Alpha-recombination is based on the reaction of Lutay et al.[38] (e.g. Figure 5) in which two RNA strands are held by Watson-Crick pairing over a splint molecule and can undergo specific cleavage and ligation. The general features of the algorithm are flexible, but to ensure realistic simulations, we required several strict criteria. First, the splint strand had to be a minimum of eight nucleotides long. The first strand to bind (hereafter "primary" strand) had to be a minimum of four nucleotides long if it had a cyclic phosphate at its 3′ terminus, or five nucleotides if it did not have a cyclic phosphate. The leaving group is any region of the primary strand displaced by the second

strand (hereafter “secondary” strand) that does not bind with any complementarity to the splint.

We required the recombination junction to be flanked on either side by at least four consecutive Watson-Crick base pairs, with a total of at least 8 consecutive Watson-Crick pairs in the entire complex. The primary strand must bind to the splint with at least four base pairs. Once there is no Watson-Crick pair, the 3′ end cannot have any further Watson-Crick complementarity to the splint. If the primary strand contains a cyclic phosphate, no unbound 3′ region is required; without the cyclic phosphate, there must be at least one unpaired nucleotide at the 3′ end. The nucleotide of the secondary strand containing the 5′ nucleophile must bind to the nucleotide adjacent to the last WC pair of the primary strand, displacing the non-WC 3′ tail with at least four consecutive WC pairs of its own.

Binding of the primary strand to the splint can result in exposed template on either side of the splint. If the exposed template is upstream of the primary strand, the secondary strand becomes the strand that is attacked by the primary strand (the “acceptor strand”) and must bind to the exposed template with at least one nucleotide in the leaving group or a cyclic phosphate on its 3′ terminal. However, if the exposed template is downstream, the secondary strand is the “nucleophilic strand”, and its 5′ hydroxyl must be adjacent to the last Watson-Crick pair of the primary strand and splint, which must have either a leaving group or a cyclic phosphate. The general scheme for random recombination is illustrated in Figure 61. Examples of alpha-recombination structures generated at random by our program are shown in Figure 62.

5.21 Activity of alpha-reactions

In order to examine the effects of alpha-recombination on a pool of random RNA we characterized three activities resulting from an alpha-setup, cleavage, ligation, and background cleavage. Cleavage refers to the specific cleavage of the extruded 3′ tail of the acceptor strand by attack of the 2′-OH of the last bound nucleotide on the adjacent 3′ phosphodiester group to produce a cyclic phosphate. Ligation refers to the attack of the 5′-OH of the nucleophilic strand on the cyclic phosphate created by the cleavage reaction to form a new phosphodiester bond. Finally, background cleavage is non-templated, spontaneous cleavage of any phosphodiester bond in any molecule that generates a cyclic phosphate on the terminal of the leaving 5′ strand.

We assigned probabilities for each activity manifested in the alpha setup. Initially, splint-catalyzed cleavage is assigned a probability of 0.3 for each alpha setup, ligation is given a probability of 0.2, and background cleavage, which can occur in any strand, has a probability of 0.01. The specific cleavage of the splint is the most probable of the three activities, while background cleavage is lower than both the splint-catalyzed cleavage and ligation. Under these conditions, a full recombination, which is cleavage followed immediately by ligation, has a net probability of 0.06, whereas the individual probability of only cleavage or ligation (to a cyclic phosphate) is much higher.

5.22 Oligomer distributions

A pool of randomized oligomers was created with a negative exponential distribution according to the formula $N_L = M / (2^{L-1} * L)$ where M is the number of monomers, L is the length of oligomers, and N is the number of oligomers (of length L). Typically, using M = 3,200,000, the maximum length is 17 nucleotides long, with

3,200,000 monomers, 800,000 dimers, 266,667 trimers, 100,000 tetramers, 40,000 pentamers and so on. The pool was a vector of RNAStrand objects that were created for each size class $L \geq 5$. Oligomers smaller than 5 nucleotides long were omitted for computational considerations and the fact that such oligomers are too small to react by our strict model of alpha-recombination. For pools with a flat distribution of 16mers, the initial pool contained exactly 50,000 random 16mers.

Complete simulations of recombination were carried out over 0.1 to 20 billion generations. In each generation, three strands were chosen at random. Firstly, a splint strand was chosen to serve as a template. Secondly, a random strand was chosen to be the primary strand and aligned 5′-3′ over the 3′-5′ splint strand in every possible way to find double-stranded conformations that met all requirements for alpha recombination. If the primary strand could bind in any such way, a second random strand was chosen to be the secondary strand and tested to see if it could pair with the exposed template before or after the hydrogen-bonded region of the first strand and splint. If exposed template was downstream, the third strand's first four 5′ nucleotides had to bind directly adjacent to the hydrogen-bonded region to position the 5′-OH for attack. If the exposed template was upstream of the hydrogen-bonded region, the strand could be aligned in any conformation with four bound nucleotides directly adjacent to the hydrogen-bonded region that left at least one nucleotide in the 3′ tail overhanging the alpha structure, or with a 3′ cyclic phosphate terminal adjacent to the 5′-OH of the first strand. Using these strict criteria with our chosen exponential distribution, three strands chosen at random from a random pool form an alpha-recombination setup approximately once every 100,000 tries for activated pools and three in every million tries for unactivated pools. This nuance is not

to be ignored; a pool that has higher rates of alpha setup formation will have higher net recombination even if the chemical rate is the same.

5.23 Mechanistic details of recombination

Upon occurrence of a successful alpha-setup, the extruded tail was specifically cleaved with probability P_{cleave}, with typical value of 0.3. If the acceptor strand had no tail but was terminated with a cyclic phosphate, the cleavage step was skipped. If cleavage was successful, the tail was removed and the 3′ terminal of the acceptor strand gained a cyclic phosphate, which was assumed due to cleavage. If the tail was less than 5 nucleotides long, it was too small to be a substrate or splint and was subsequently flagged to avoid being chosen, though it was counted towards the final distribution for computational reasons (the rate of alpha setup formation is independent of the number of strands but the simulation can be slowed by quantities of inert substrates).

For ligation, the acceptor strand, now terminated with a 2′-3′ cyclic phosphate, could be attacked by the 5′-OH of the nucleophilic strand with probability $P_{ligate} = 0.2$. If strands formed an alpha setup but failed to cleave, the entire complex of 3 strands dissociated back into the pool. If the acceptor strand cleaved but the nucleophilic strand failed to ligate, the complex dissociated back into the pool as three strands; the acceptor now being a shorter strand terminated with a cyclic phosphate and the tail of the acceptor having been previously cleaved and in solution as a separate strand. If both cleavage and ligation are successful, the acceptor and nucleophilic strand become one strand, and the double stranded complex dissociates.

The final step of the simulation is background cleavage. As the alpha setup occurs at most approximately every 1/100,000 generations, on every 100,000th generation (independent of alpha setup formation) a strand is chosen at random from the pool and undergoes cleavage at a random bond with probability 0.01. If the strand cleaves, a random bond is subsequently chosen and the strand is broken into two strands at that point, leaving behind a cyclic phosphate on the 5′ piece and a 5′-OH on the 3′ piece, which also serves to activate the 5′ piece at the expense of making it shorter. The probability of spontaneous cleavage at a random bond is necessarily and justifiably lower than both the cleavage and ligation reactions catalyzed by splinting because spontaneous cleavage is uncatalyzed. In this simulation, the baseline rate of background cleavage is 1/30th the rate of splint-catalyzed cleavage.

5.24 Structural feedback

For simulations with structural feedback, a simple folding algorithm designed to match upstream or downstream base pairs was created to identify four specific types of hairpin structures, tetraloops, pentaloops, hexaloops, and heptaloops. The algorithm is performed on every strand ≥ 14 nt resulting from a successful cleavage and ligation, background cleavage, or cleavage-only event. The algorithm iterates over every possible hairpin that could result from direct alignment of upstream or downstream base-pairing, ignoring possibilities of bulges due to computational considerations (Figure 63).

In order to make a stable hairpin, our model requires a loop of 4-8 nucleotides. The closing base pair must be G-C; if it is A-U it must have a G-C pair immediately adjacent and downstream. The loop must be closed with at least 3 consecutive Watson-Crick pairs and there must be a minimum of 10 hydrogen bonds in the closing helix to

qualify as a hairpin. To meet these requirements (in terms of hydrogen bonds), the number of possible hairpins of a given loop size S for oligomer of length N is (N – 10 – S). Thus, a 20mer has 6 possible tetraloops, 5 possible pentaloops, 4 possible hexaloops, and 3 possible heptaloops. The program iterates over all possible hairpins to find the hairpin with the lowest free energy, which is used in the simulation. If no structure meets the requirements, the oligomer is considered unstructured. If the molecule is a hairpin, it is given a 20% chance of having ribozyme activity, which is characterized by the specific features of the hairpin.

We designed feedback mechanisms for hairpins with ribozyme activity. A generic tetraloop, pentaloop, or hexaloop can accelerate specific cleavage, altering the overall probability of cleavage according to a linear formula $P_{cleave} = P_{cleave} + 0.01 * ([te] + [pe] + [he])$ where [te] is the tetraloop count, [pe] is the pentaloop count and [he] is the hexaloop count. Heptaloops, and specific tetraloops of the form GNRA and UNCG can accelerate the ligation step of alpha-recombination according to the formula $P_{lig} = P_{lig} + 0.02 * ([GNRA] + [UNCG] + [hep])$, where [GNRA] is the number of GNRA tetraloops, [UNCG] is the number of UNCG loops, and [hep] is the number of heptaloops. Specific AANA tetraloops accelerate the rate of background cleavage, functioning as nonspecific nucleases, and specific CUUG tetraloops allow molecules to resist AANA activity by the formula $P_{bkgd} = P_{bkgd} + 0.02 * ([AANA] – [CUUG])$. Finally, GNRA and UNCG tetraloops are 10-fold resistant to catalyzed cleavage; as some of the most stable types of hairpins, they are less likely to denature into the linear strands necessary for the alpha model.

To confirm the validity of our general alpha recombination model, we ran a brief simulation with an initial pool composed of 6572 strands of 5′-CUCUCCUUCCUGAAAA and 3428 strands of 5′-GAGAGCAGGAA, the oligomers used in the 2007 reaction of Lutay et al.[38] Allowing this reaction to run iteratively until all of the 16mers have reacted produces a distribution of recombinant strands with the predominant products being the cleavage and/or ligation products of the alpha reaction, confirming that our simulated mechanism is accurate and correct (Table 3). None of the products resulting from this reaction have any significant structure; no alteration of the cleavage or ligation rates is observed in any simulation of this test distribution. While this is not exactly a true representation of what would happen *in vitro*, where a majority of the starting material would remain unreacted, this simulation does have the property that a large majority of the products are the cleavage product CUCUCCUUCCUG, and the next largest product is the true recombination product, while there is only a small amount of larger recombination products. Thus, the result is conceptually accurate in terms of the proportions of recombinant products.

5.25 Replenishment and exchange

For simulations with the exponential distribution in which the pool periodically exchanges strands with its environment, each strand in the pool 10 or fewer nucleotides long had a 20% chance to be removed during replenishment. Strands larger than 10 nucleotides had a progressively decreasing chance to be removed by one percentage point per nucleotide up to 25 nucleotides; all strands greater than 25 nucleotides long were capped at a 5% chance of removal. This reflects the possibility of larger recombination products being adsorbed to mineral surfaces or retained by environmental gradients. In

addition, all the unusable monomers, dimers, trimers, and tetramers were assumed removed and replaced, since they would be less likely to adsorb to surfaces and also more likely to escape gradients. The total count of removed strands and nucleotides was summed as nucleotides, and then this sum was subject to the exponential distribution, giving a number of strands of each size that should be added. Random strands in these amounts were added to the pool; the net effect is so the overall concentration is kept relatively constant throughout the simulation. For exponential distributions, replenishment happens every 500 million generations in our simulation, thus for a 10 billion generation simulation, there are a total of 20 replenishment/exchange events.

5.3 Simulation results

To begin with, we examined how the length distribution of random pools of abiotically generated RNA molecules changes with time. This simulation is similar to our previously published result – there are no structural effects – but here we more carefully examine the effects of cyclic phosphate preactivation. Using our standard parameters, we began with 50,000 randomized 16mers with and without preactivated 3′ terminals and reacted them for 100 million generations (Figure 64). In this scenario, both populations undergo a significant redistribution to produce an abundance of new strands, the majority of which are in the range of 4-12 nucleotides. However, there is also a significant quantity of new strands longer than 16 nucleotides, whose sum ranges from 4-12% of the total starting material. We also find in this simulation that the preactivated pool has considerably improved length expansion compared to the unactivated pool, which is consistent with the *in vitro* results of Mutschler et al.[51]

We computed the average number of alpha setups formed in each type of simulation over five simulations and found it was 37541 for the unactivated pool and 40673 for the activated pool. In addition, we computed the average total number of cleavages, ligations, and recombinations in each pool (Table 4). Here, recombination refers to events where a cleavage event is immediately followed by a ligation, giving the effect of one-step recombination versus one-step cleavage or ligation. Each pool has a similar number of cleavages but the preactivated pool has, perhaps unsurprisingly, a substantial increase in the number of ligations. Ligations can still happen in the unactivated pool because cleavage initiated by a 2′ hydroxyl will leave a cyclic phosphate on the reactant strand – breaking a strand into two strands also serves to activate one of them – but ligations can only happen after a cleavage event has activated a strand.

In addition, the preactivated pool has a higher average number of both cleavages and one-step recombinations. In the absence of any rate increase, this reflects the fact that the preactivated pool had more total alpha setups, and thus had more chances to recombine. This result indicates that even if the recombination rate is unchanged, the rate of recombination per strand can be passively effected by the composition of the solution – any activating agents, stabilizing agents, or reduction of solution heterogeneity could result in a net increase in recombination because of the increased rate of alpha setup formation that is independent of the chemical reaction rate. We noted in the Methods that the activated pool forms alpha setups at a rate of approximately 1/100,000, whereas unactivated oligomers only form alpha setups a rate of 3 in a million; this distinction is now seen to produce a passive net increase in overall recombination in the activated pool.

We next explored altering the actual rates of recombination by doubling the rate of cleavage from 0.3 to 0.6 while leaving the ligation rate unchanged. This might represent conditions with a more basic pH where cleavage is expected to be more rapid but ligation can still occur. In this scenario, the unactivated pool now has an improved distribution relative to the activated pool because of the increase in activation due to cleavage (Figure 65). If we increase the cleavage rate to 0.8 and reduce the ligation rate to 0.1, the effect is even more pronounced (Figure 66). However, if we increase the ligation rate to 0.35 from 0.2 but leave the cleavage rate unchanged, which might represent cold eutectic conditions or alternative solvents that favor ligation, the preactivated pool produces a distribution whose length expansion is greater than the comparable preactivated simulation with unaltered ligation rates (Figure 67).

5.4 Structural feedback by ribozyme catalysis

In many origin-of-life scenarios for the RNA World, a key component for bootstrapping inactive pools of strands into a lifelike state is the activity of ribozymes, typically polymerase ribozymes, that are able to replicate themselves and other strands. Recombinase ribozymes are generally small ribozymes that have transient effects on the solution as a whole. The effects of a single ribozyme in our solution is negligible, but the accumulation of many ribozymes will have a positive feedback effect on cleavage, ligation, and background cleavage.

When structural feedback is added to our flat distribution of 16mers for both the unactivated and activated distribution over 100 million generations, it can be observed that both distributions are substantially enhanced compared to simulations without structural feedback (Figure 68). In the preactivated pool, strands as large as 162

nucleotides can be observed, a size on order of that required to form an RNA polymerase ribozyme. Likewise, in the unactivated pool, the length expansion reaches 96mers, far above any of the simulations without recombinase feedback. Of note in both distributions with structural feedback is the fact that monomers, dimers, trimers, and tetramers are created in the greatest amounts compared to any other size over the same generational time frame.

5.5 Recombination of an exponential distribution

To make the simulation more prebiotically appropriate, we applied the algorithm to the exponential distribution described in the Methods, and increased the total number of generations to 10 billion. In this case, a large fraction of the total strands are much shorter; a quorum are exactly five nucleotides long, which we have designated the minimal size for alpha-recombination. The increase in number of generations is necessary because the size distribution makes the rate of alpha setup formation less likely than in the flat distributions; only a relatively small minority of strands even have the length required to be a splint. We again computed the averages for 10 simulations of activated and unactivated oligomers in an exponential distribution. In this scenario, the average number of alpha setups over 10 billion generations is 30,597 for the unactivated pool and 28,203 for the activated pool.

Next, we examined the results of an unactivated distribution with a total of 69,469 oligomers in the negative exponential distribution ranging from 5-17 nucleotides over 10 billion generations and find that there is a modest distributional shift that depletes the intermediate sizes (5-8 nt) and increases the formation of oligomers from 9-19 nucleotides (Figure 69). Likewise, a natural consequence of this redistribution is the

increase in amounts of monomers, dimers, trimers, and tetramers, which correspond to the small leaving groups in the alpha reaction and the results of background cleavage. In contrast, the preactivated pool again produces a redistribution with superior length expansion compared to the unactivated pool (Figure 70).

We repeated our simulations with increased cleavage (Figure 71), increased cleavage and lower ligation (Figure 72), and only increased ligation (Figure 73). The simulations with increased cleavage both mirror the results of the flat distribution; the preactivated pools are less efficient and the unactivated pools show some improvement. When only the ligation is increased, the unactivated pool again shows modest improvement. However, under conditions of increased ligation for the preactivated pool, the phenomenon occurs that the entire distribution recombines so much that it eventually decays itself almost entirely into nucleotides and oligomers of sizes four or less. This demonstrates one of the limitations of a closed system – without some kind of replenishment and interaction, or without the ability to ligate unactivated monomers, alpha recombination over long time periods will ultimately reduce the entire pool into very short fragments. We examined the fate of this preactivated distribution at each 2 billion generation timepoint, including the peak of the distribution at approximately 9.6 billion generations, in which the maximum oligomer size of 187 nucleotides emerges (Figure 74). This graph reveals a more profound detail of recombination over long time periods – before the total reduction of oligomers into nucleotides, there is a systematic length expansion of the entire distribution, and a peak which is followed by collapse.

Finally, we repeated our simulations with structural feedback on the exponential distribution (Figure 75). Here, the results for the exponential distribution are fairly

modest in comparison to the flat distribution of – 10 billion generations are not long enough to convert the short strands into strands long enough to be structured. Nevertheless, over the course of the simulation, the rate of cleavage in the preactivated pool rises from 0.3 to 0.41 and from 0.2 to 0.24. While these numbers are not as high as our artificially boosted scenarios, including the ligation scenario where the distribution collapsed, they do demonstrate that structural feedback can accelerate recombination even in a distribution of oligomers that are mostly too small to have significant structure.

5.6 Replenishment scenarios

In all of the above scenarios, the algorithm is performed on a static pool that does not change with time and is isolated from the environment. If all of these simulations are run for a very long number of generations, or to near infinity, the entire pool would eventually decay into monomers – the combination of maximal alpha recombination and background cleavage – as evidenced by the case of the exponential distribution with artificially boosted ligation. In the short term, this is a realistic simulation of a laboratory reaction, in which the reaction is often closed, but does not necessarily reflect what may have happened on the primordial earth. Although it is certainly possible and plausible that isolated systems could have arisen and provided the conditions for unhindered sequence expansion and recombination, any pool of recombining oligomers would have ultimately needed a source of further interaction or replenishment to become life.

One essential feature that might be necessary for a pool of RNA to achieve a steady state of recombination and length expansion is the recycling of monomers. This can be addressed in many legitimate ways, such as the presence of activating agents and catalysts to re-ligate activated monomers back into sizes appropriate for recombination.

We specifically chose to address this problem by adding in a feature of replenishment, in which the entire pool periodically exchanges some of its strands with the environment. An example to describe this scenario might be that of a tide pool, in which recombination occurs during low tide, but during high tide, new material dispersed from sea-floor catalysts or activating surface chemistry is exchanged for unactivated monomers and small recombination byproducts, as well as some of the existing strands, so that the pool is refreshed by an injection of new oligomers.

We have assumed that the feedstock for our algorithm is generated by an abiotic polymerization process; we now assume that there is some dynamic combination of recycling process and/or environmental flow that brings fresh oligomers into the pool to replace the unusable monomers and other short pieces. One question in this above scenario is whether the longer products of recombination would be retained in such a circumstance. To justify our feature of replenishment, we will assume that recombinant products in our pool may adsorb to a mineral surface in the pool, allowing retention of longer products. An alternative scenario could involve RNA oligomers trapped in a lipid vesicle that is able to release nucleotides and small RNAs while occasionally growing by merging with other vesicles. This exchange of material is set to happen every 500 million generations in our pool. In theory, there could be many alternative scenarios.

With replenishment added in to our simulation, time should not be a significant factor as the simulation should not be able to completely decay; in addition, the occasional loss of strands will hinder structural feedback. Therefore, we increased the distribution to begin with M = 4,000,000 instead of 3,200,000, and ran unactivated and preactivated pools with and without structural effects over 10, 20, and 30 billion

generations. In the time frame of 10 and 20 billion generations, no significant changes are observed for either unactivated or preactivated pools. In addition, over 30 billion generations, there is no significant change for the unactivated distribution. However, for the preactivated pool over 30 billion generations, the distribution gradually rises, accumulating both sequence and structure until there is a sudden and massive spike in the number of structures, as well as the catalysis rate, that brings the collective rate enhancement to its maximum. The result shows a rough steady state distribution of oligomers ranging up to 200 nucleotides long (Figure 76).

A close examination of the general features of the replenished simulation reveal that the steady state distribution only reaches its zenith after 29 billion generations (Figure 77). For most of the simulation, the distribution of the pool has the same general feature as the corresponding preactivated pool without structure, and the rapid length expansion of oligomers occurs suddenly in a very short time frame. Analysis of the structural content (Figure 78) and types of structure present in the pool (Figure 79) also reveal the same general phenomenon. However, a plot of the catalysis rates versus generation shows that the underlying rates of cleavage and ligation begin to increase early in the simulation due to ribozyme accumulation (Figure 80). While these rates also feature a similar spike towards the end, it is clear that one of the drivers for this spike must be the gradual rise of the recombination rate – or more specifically, the rates of cleavage and ligation – the component reactions of alpha recombination.

As a final analysis of our replenished simulation, we examined the general formation of alpha setups over time, and find a rather remarkable result – even though the rate of formation of alpha setups is completely independent of the catalysis rates for

cleavage and ligation, the amount of alpha setups being formed near the end of the simulation is dramatically accelerated (Figure 81). Since alpha setups are formed by the choice of random strands, it is unaffected by the recombination rate, and is also not substantially affected by the size of the pool – whether the pool is small or large does not change the fundamental odds that three random strands should form an alpha setup – unless there is an underlying lack of diversity or if the pool is prohibitively small, but this is not the case here.

Ultimately, the most likely reason for the rise in alpha setup formation is because by the end of the replenished simulation, many of the strands in the pool have acquired structure, and a majority have already undergone some form of recombination. This result strongly suggests that oligomers that have previously recombined are substantially more likely to undergo further recombination, a consequence of their established secondary structures and innate ability to bind to other strands in the pool. In our simulation, the sudden spike in alpha setup formation, along with the spike in structural formation occurs only at the end of the simulation, suggesting that a large proportion of the pool needs to have undergone recombination to make the spike possible.

Therefore, in a pool of short RNA oligomers undergoing recombination, if the pool is able to exchange strands with its environment, periodically replenish itself with fresh material, and is able to retain a significant proportion of its recombinant strands, we conclude that there is a fundamental tendency of the pool to undergo increasingly accelerated recombination, leading to both substantial length expansion of the pool, and the substantial accumulation of structured RNA species whose sizes are on the order of those needed to make advanced catalysts.

5.7 Discussion

Previous models of RNA recombination have generally treated recombination as a one-step, fully reversible process. Nevertheless, the features of the alpha reaction, cleavage of an extruded tail and ligation over a splint with Watson-Crick pairing, give the reaction some degree of irreversibility. Although the cleavage and ligation reactions are themselves reversible, in order to reverse the reaction fully, another template is required to splint the first half of the molecule and provide Watson-Crick pairing for the previously lost tail to displace the new oligomer at the exact bond previously ligated. Furthermore, alpha recombination is not necessarily hindered by misplaced pairs or bulges, and other mechanisms of recombination are possible. Recombinant products of one alpha recombination might easily react with new molecules different than the molecules required to reverse the reaction, and thus with time, this "scrambling" is unlikely to be reversed by the exact scenarios required to reverse it. This is one critical feature of recombination that has been previously ignored by various RNA world simulations – although the recombination events themselves are energetically neutral and reversible, the resulting genetic scrambling is highly irreversible, and the sequence space exploration wrought by recombination events could give rise to molecules with real catalytic effects.

The second key point to the irreversibility of alpha recombination is the base rate of cleavage among RNAs. In theory, any 2′-OH should be able to initiate a bond-breaking cleavage reaction at some background level. Although the opposite reaction, ligation of a 5′-OH to a cyclic phosphate generated by cleavage is possible, it is far less likely than the cleavage reaction because the nucleophile and substrate are separated in solution. It is

only with the presence of a template that the cleavage reaction can be effectively reversed. With time, an isolated system of RNA molecules without additional input and without some activity to ligate nucleotides or small strands together will completely decay into monomers.

Why then, does recombination matter? It matters because prior to complete decay, a solution of random or semi-random RNAs undergoing alpha-recombination will inevitably produce longer products, which could have any number of structures and activities. They might accelerate recombination, but in the right circumstances, they might induce ligation as well. They could activate the small waste products of recombination so that they could be ligated again. They could catalyze some form of peptidyl transfer to produce polypeptides with activities of their own. They could act on sugars, carbohydrates, or even sulfur compounds to effect some form of metabolism.

It is also relevant that at the origin of life, any system of RNA molecules undergoing recombination would not be a closed system as in laboratory or computer models. Thus, the RNA that feeds recombination would have to be generated somewhere, and it is entirely plausible that the leftover nucleosides, nucleotides, or small unused pieces resulting from recombination reactions could diffuse back to their source, be recycled, re-activated, and re-polymerized by abiotic means. Indeed, it is a plausible scenario that a system of RNAs subject to regular recombination would also be periodically restocked with an injection of fresh oligomers, possibly made in part by recycled leftovers of past recombination reactions. With time, the genetic scrambling resulting from these low-energy, ceaseless recombination events could systematically expand the distribution of short oligomers to produce significant RNA catalysts, such as

an RNA polymerase ribozyme or an ancestor of the ribosome, catalyzing peptidyl transfer.

Our simulations, in which every single model alpha reaction is inherently plausible, is a starting point for further research into how recombination may generate structure and expand the lengths of prebiotically sized oligomers. We have already demonstrated that real recombination mechanisms exist for small RNAs *in vitro*. Our simulation demonstrates how a small pool of RNAs may systematically expand their lengths with energetically neutral, two-step transesterification reactions. In a closed pool, this will eventually result in complete cleavage of the oligomers, but in any environment where routine exchange and/or replenishment is possible, recombination will build up and form a steady state distribution of long oligomers that rise and fall as they are created and degraded, until the eventual emergence of an ultra-significant catalyst that could preserve or help preserve genetic information.

Chapter 5 Figures

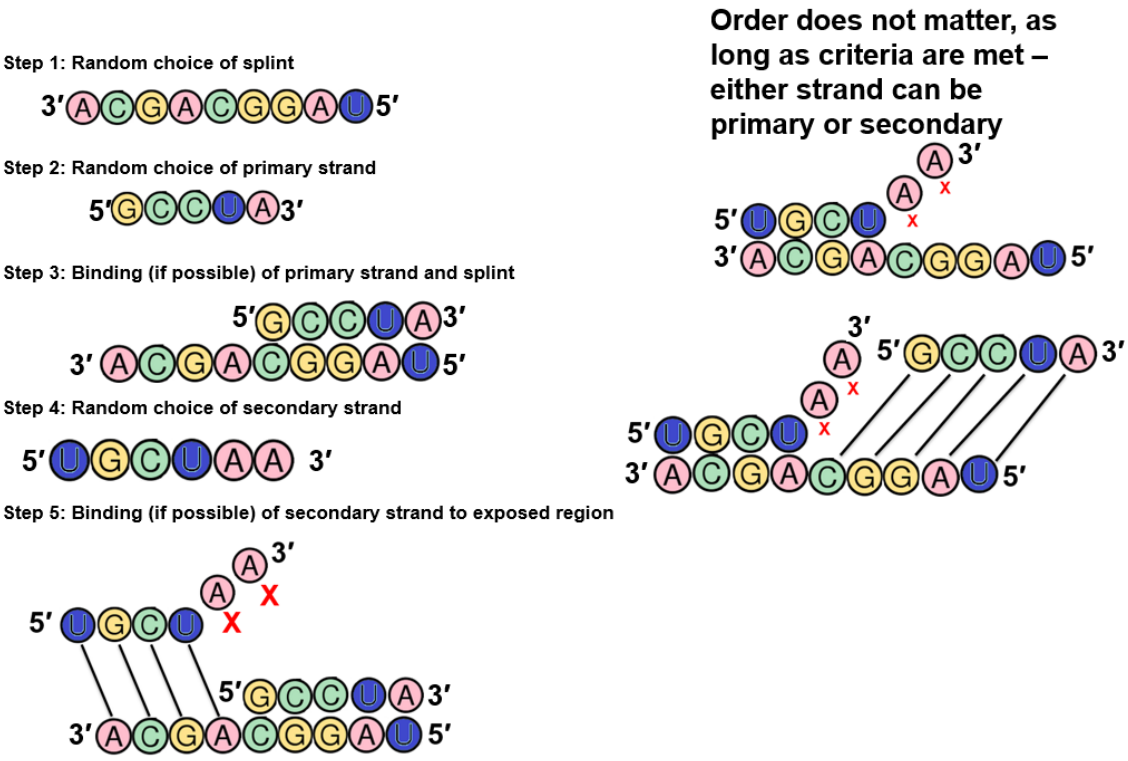

Figure 61: Schematic for random alpha recombination.

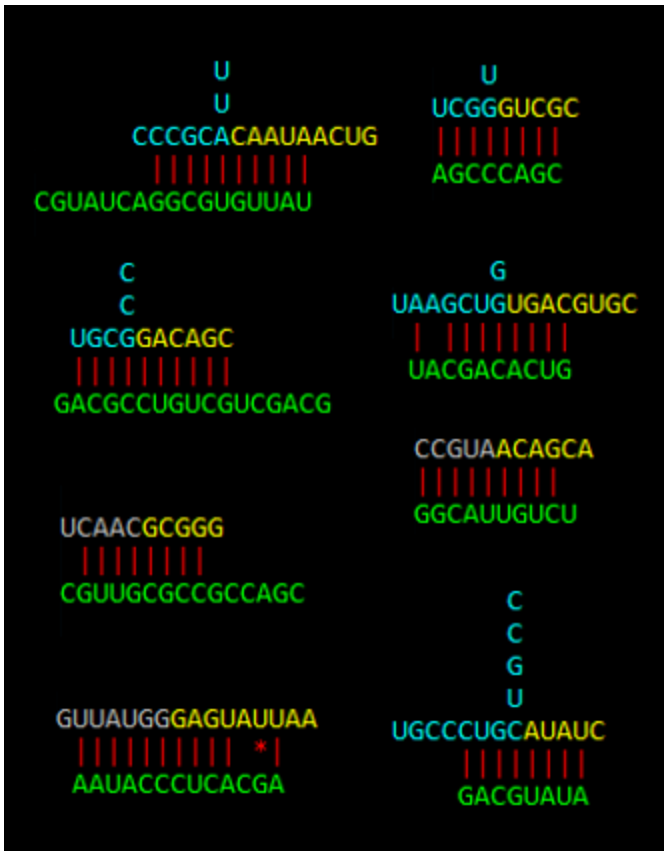

Figure 62: Examples of alpha recombination generated at random by our algorithm. The splint strand is oriented 3′-5′ in green and the nucleophilic strand is in yellow. The acceptor strand is in blue if it has an overhanging tail requiring cleavage; it is grey if there is no tail but the acceptor strand has a terminal cyclic phosphate.

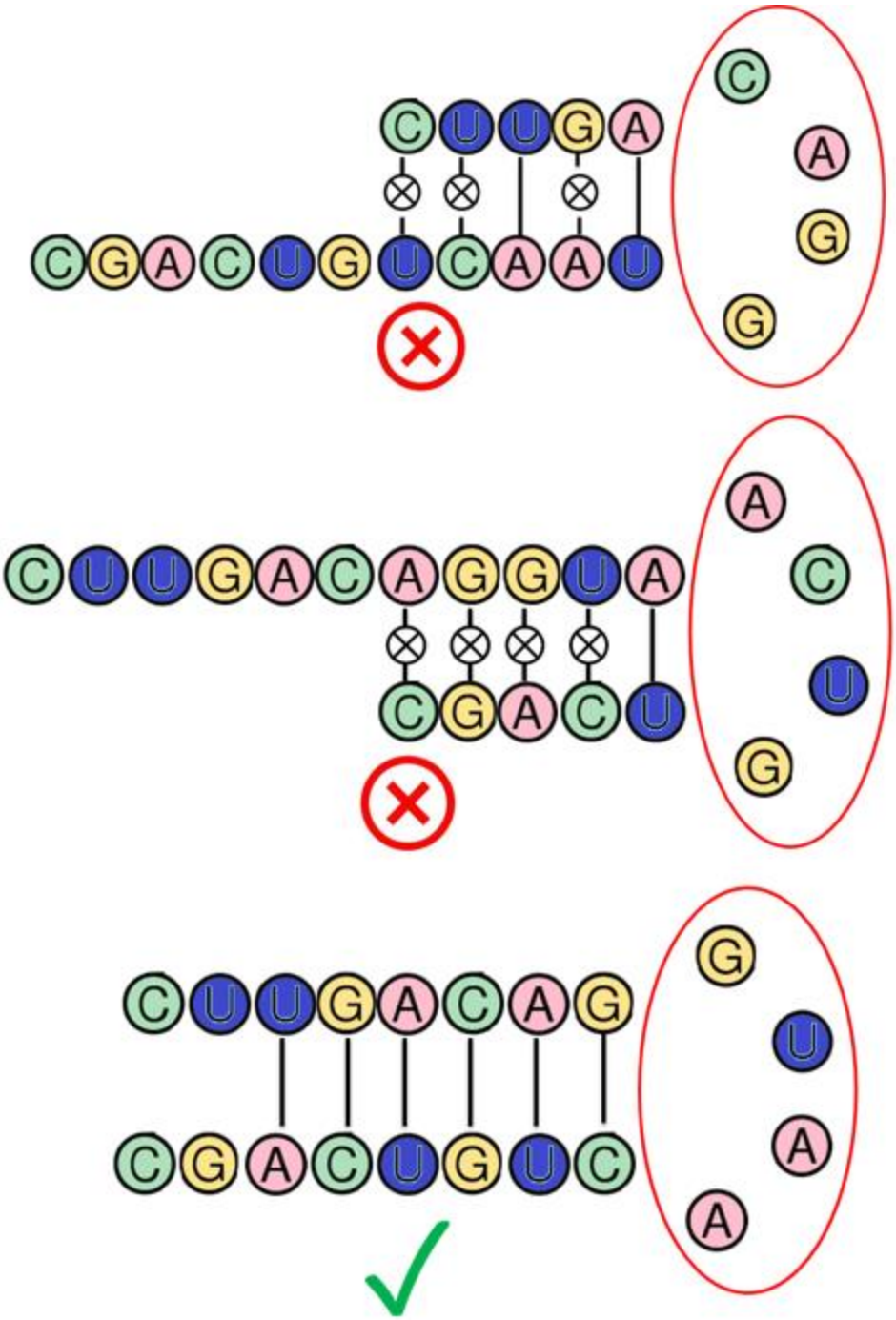

Figure 63: Scheme for finding basic hairpin structures in RNA oligomers by iteratively comparing downstream and upstream nucleotides before and after loop regions. A tetramer is used as the example, but any 4-7 nucleotides with at least five nucleotides on both of its 5′ and 3′ ends can be a loop in a hairpin loop structure.

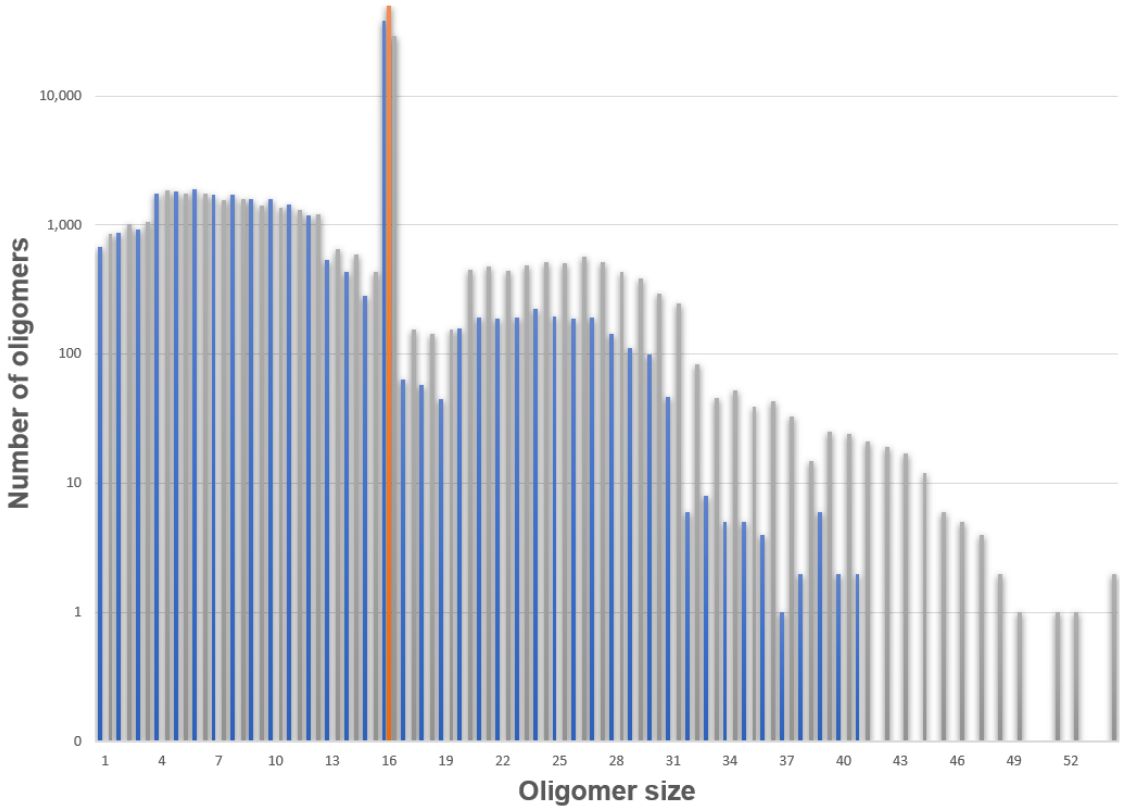

Figure 64: Length expansion and redistribution of 50,000 unactivated 16mers and 50,000 activated 16mers plotted on a log/linear scale. Each simulation was run for 100 million generations with identical parameters other than preactivation. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

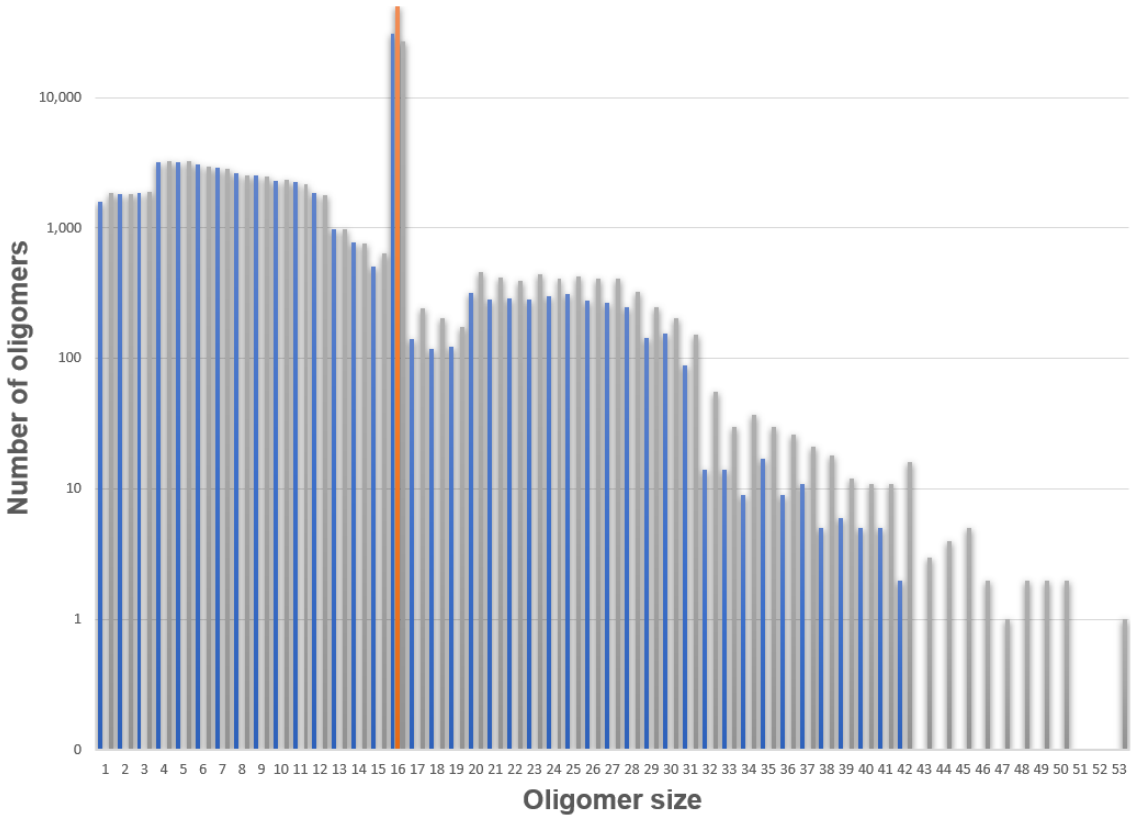

Figure 65: Repeat of simulation in figure 64, but with cleavage rate raised to 0.6 instead of 0.3. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

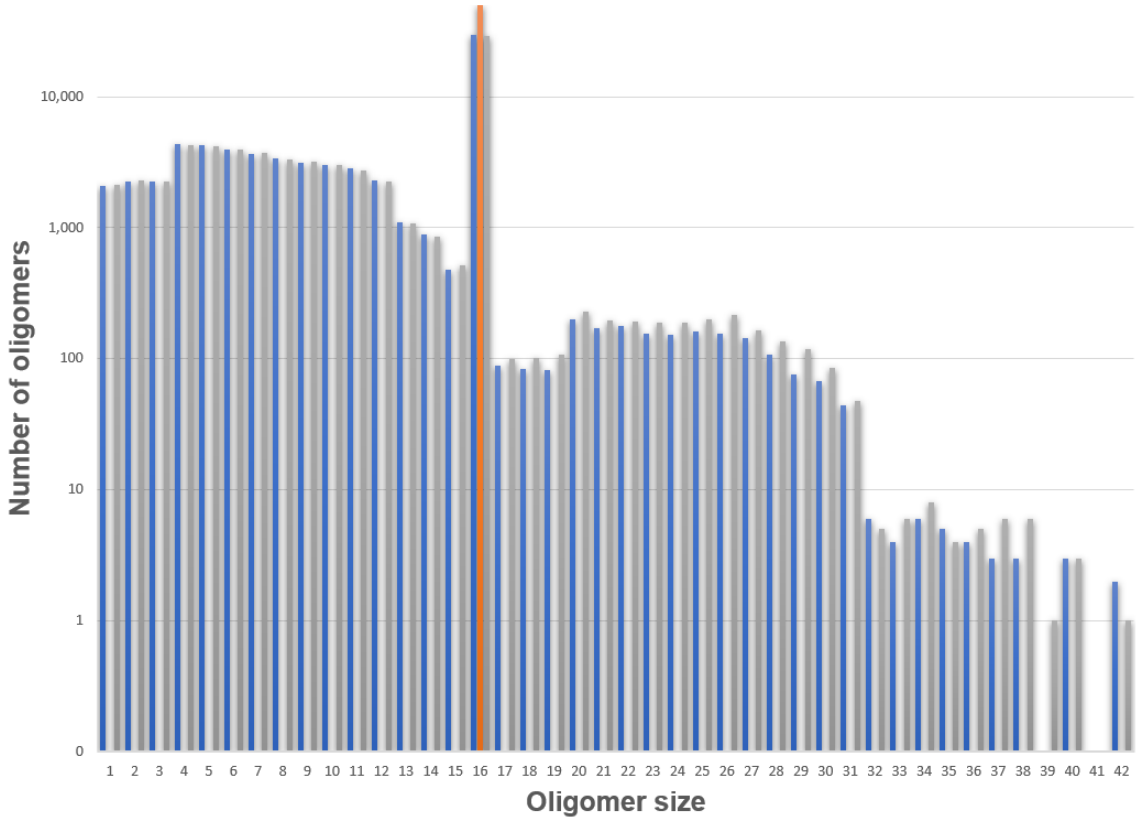

Figure 66: Recombination of the flat distribution with the cleavage rate raised to 0.8 and ligation rate reduced to 0.1. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

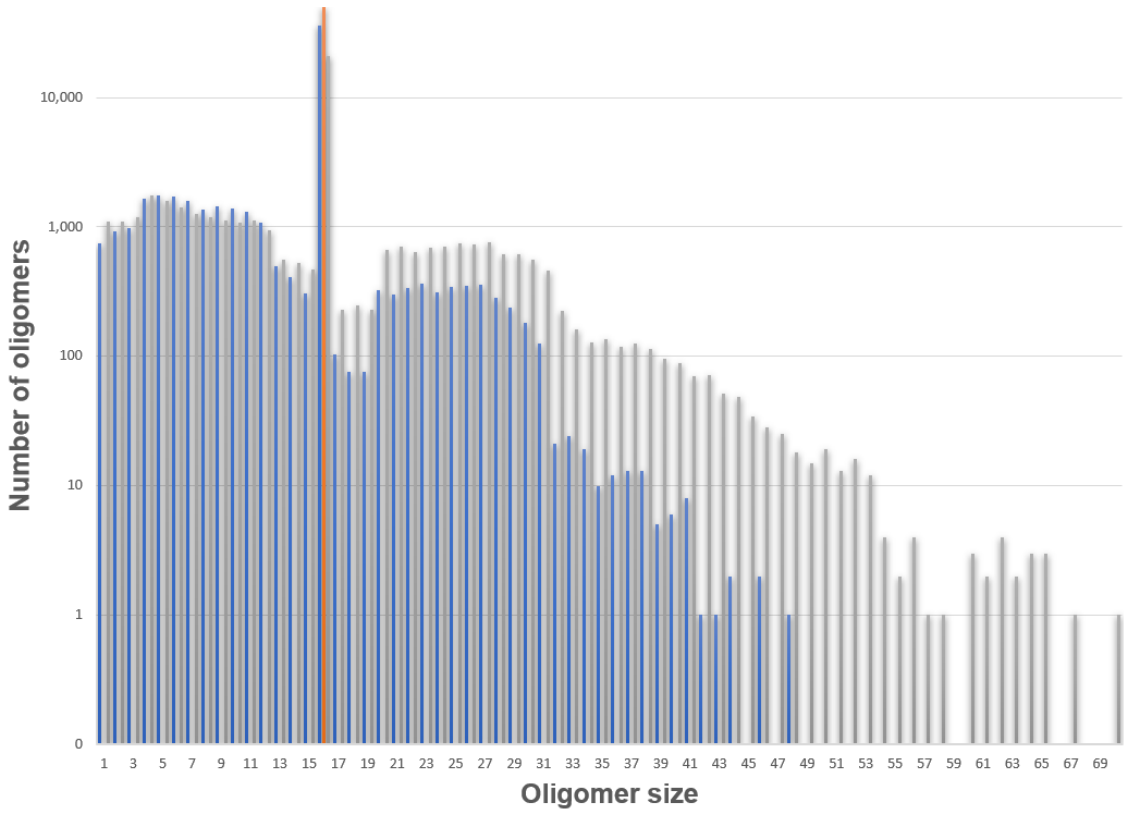

Figure 67: Length expansion and redistribution of 50,000 activated 16mers over 100 million generations with ligation rate increased from 0.2 to 0.35 and cleavage rate held at 0.3. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

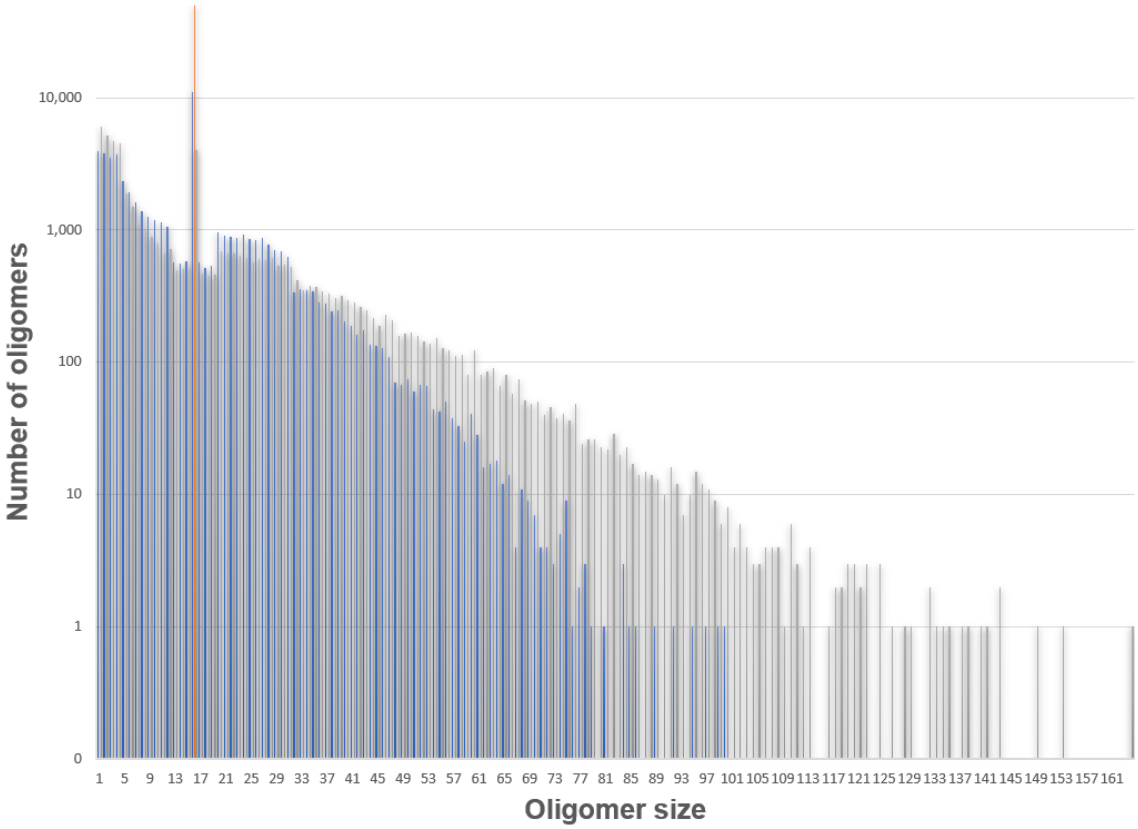

Figure 68: Length expansion and redistribution of 50,000 activated and unactivated 16mers over 100 million generations with structural emergence and ribozyme catalysis. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

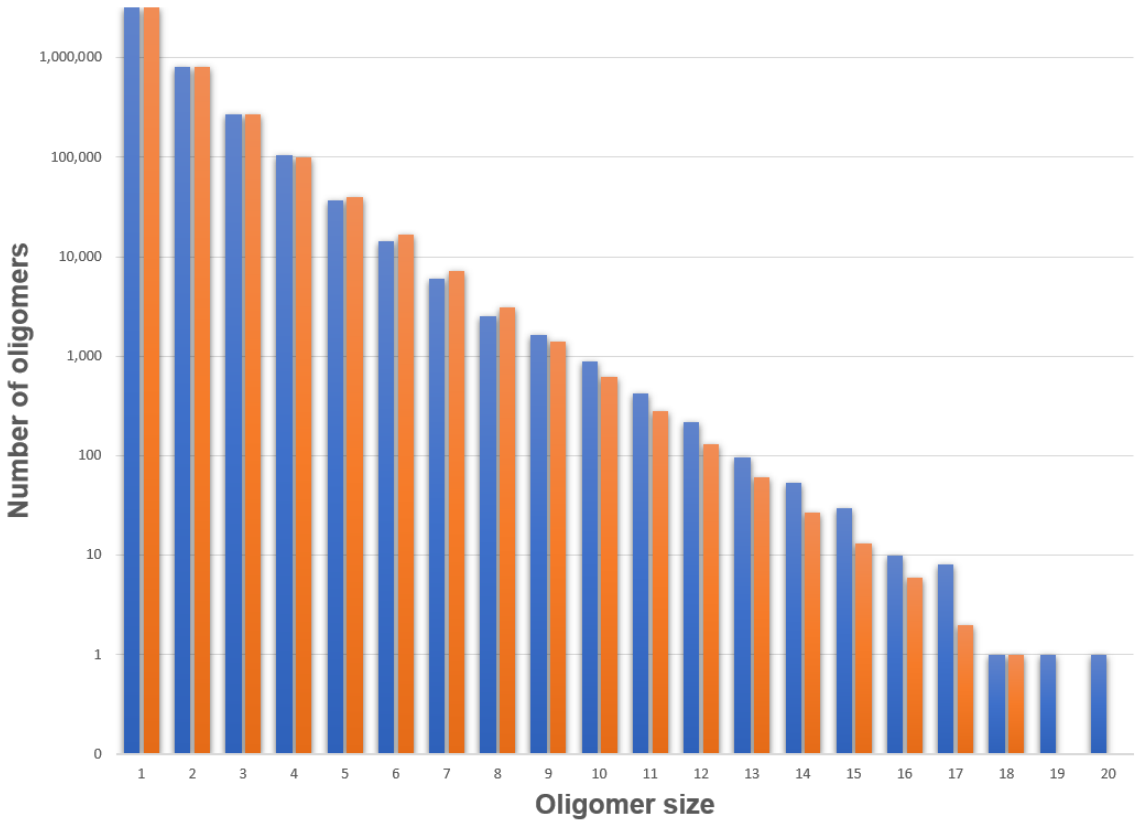

Figure 69: Modest redistribution of unactivated pool of oligomers with a negative exponential distribution from 5-17 nucleotides long undergoing alpha recombination. The initial distribution is orange and the unactivated final distribution is blue.

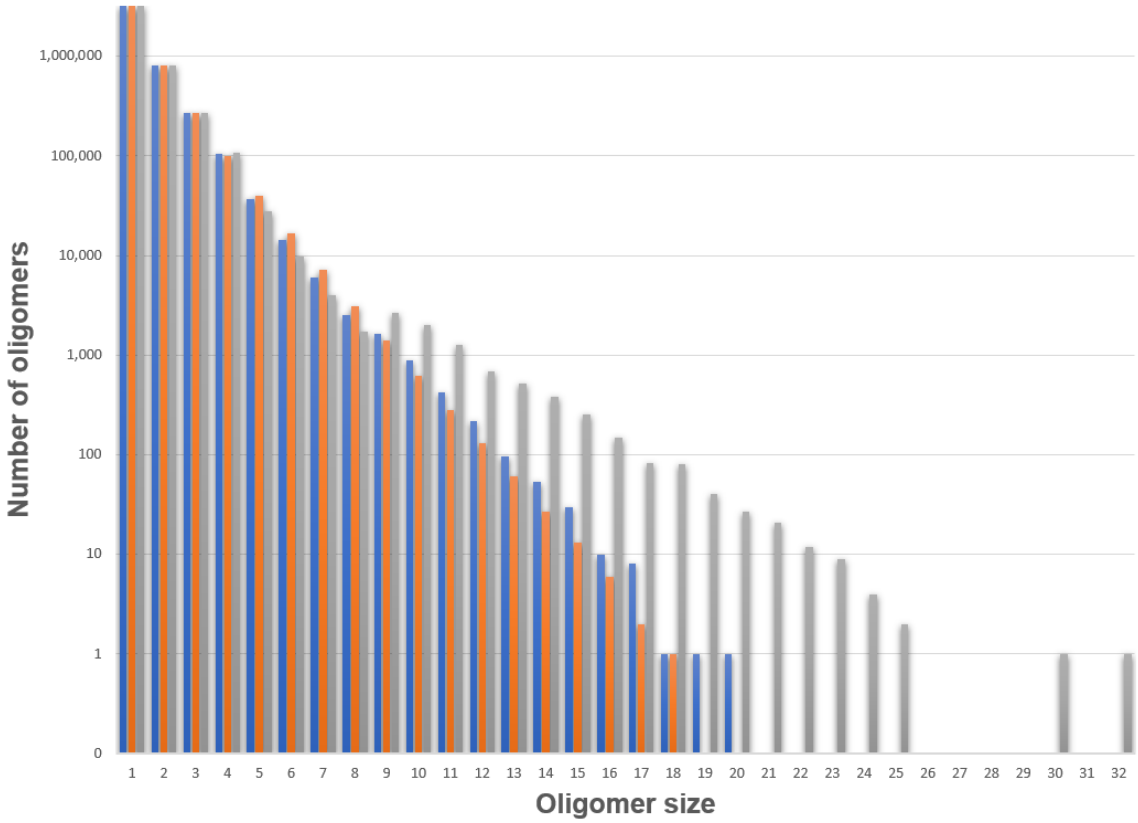

Figure 70: Preactivated and unactivated length expansion of the exponential distribution. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

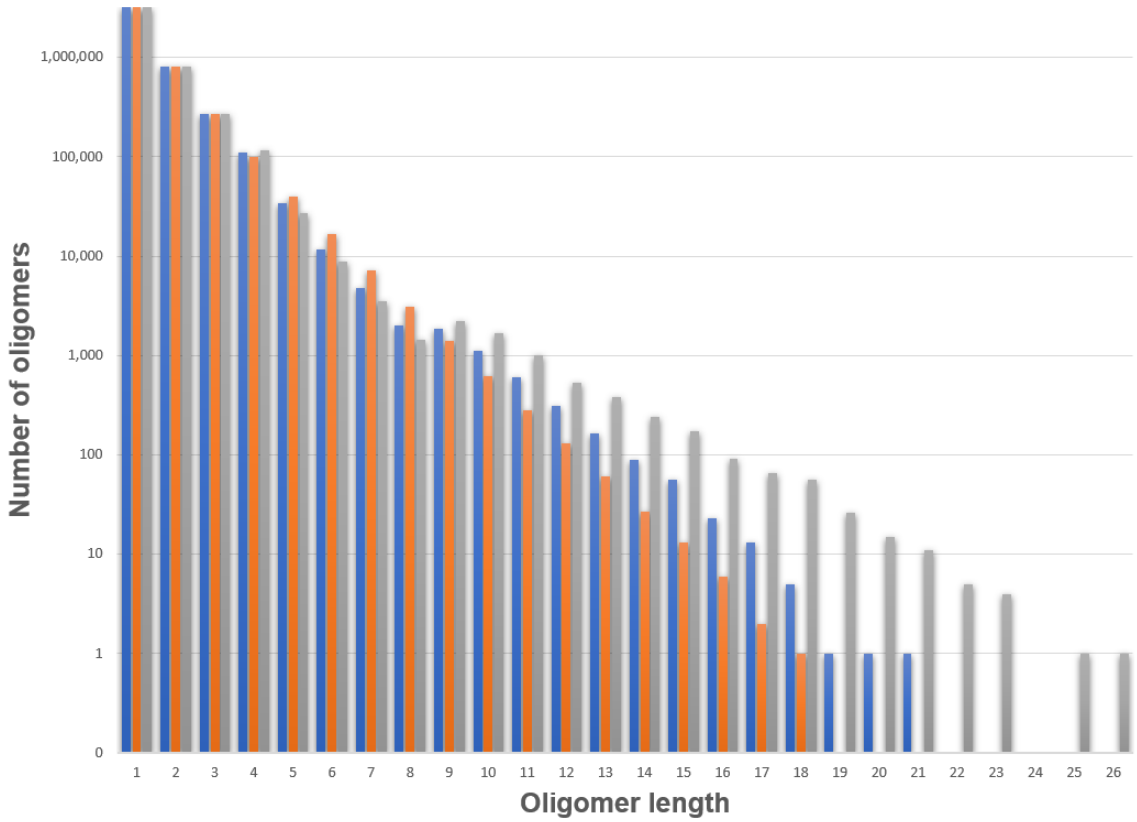

Figure 71: The exponential distribution with accelerated cleavage and unchanged ligation, with parameters identical to Figure 65. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

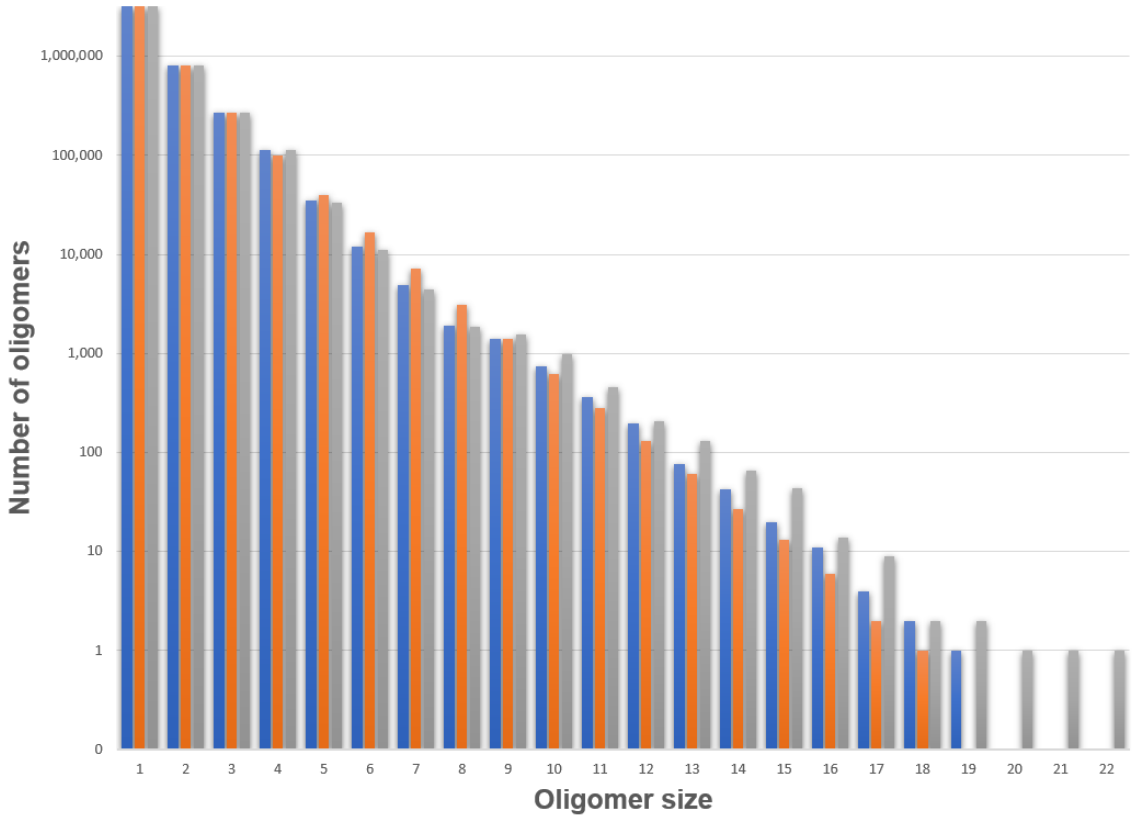

Figure 72: The exponential distribution with strongly accelerated cleavage and reduced ligation, with parameters identical to Figure 66. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

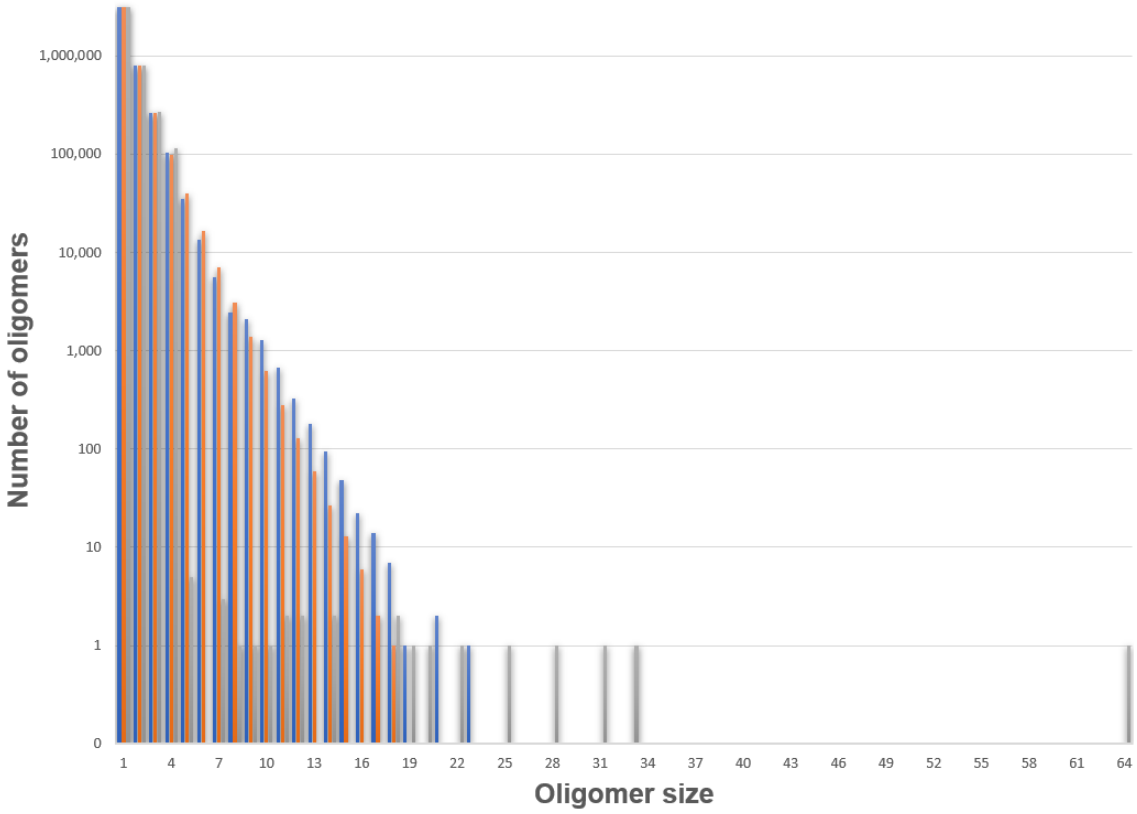

Figure 73: The exponential distribution with unaltered cleavage but increased ligation, with parameters identical to Figure 67. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

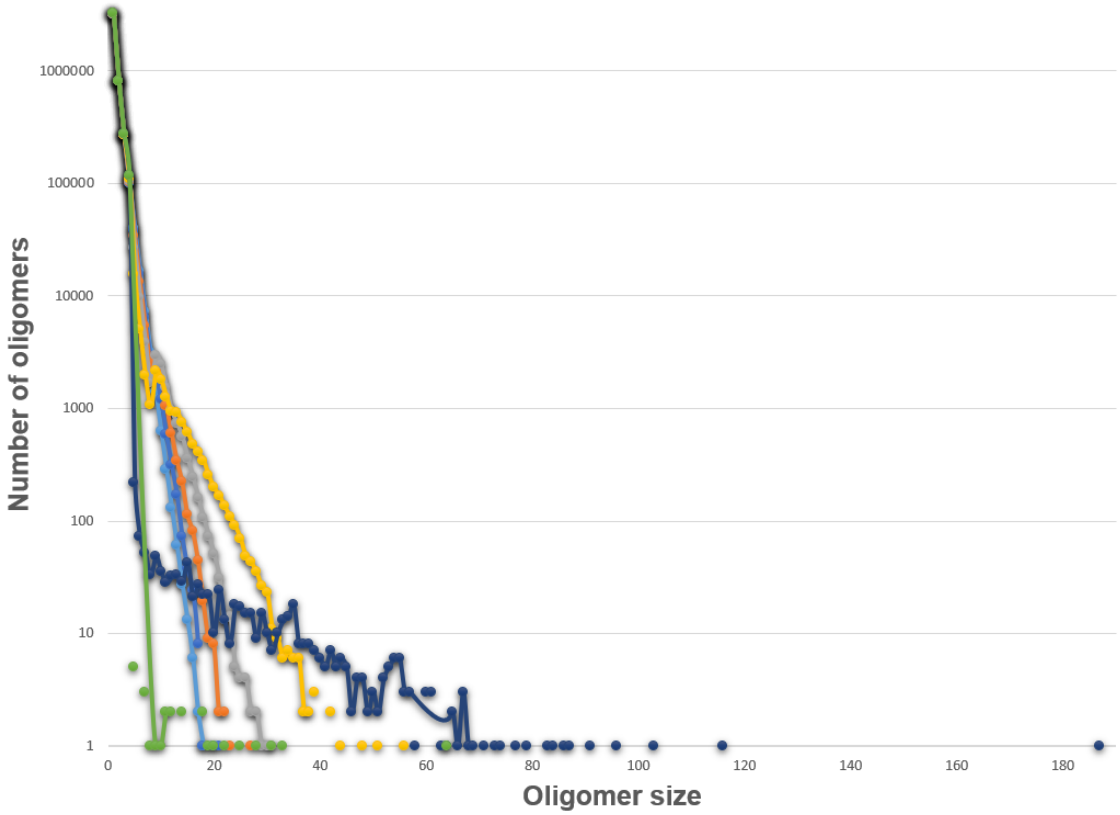

Figure 74: Examination of the preactivated distribution over timepoints of two billion generations each, showing expansion at each timepoint, the peak, and the collapse after the peak, resulting in a limited final distribution. The initial distribution is light blue, the 2 billion generation timepoint is in blue, the 4 billion timepoint is orange, the 6 billion timepoint is grey, the 8 billion timepoint is yellow, the peak distribution is dark blue, and the final distribution is in green.

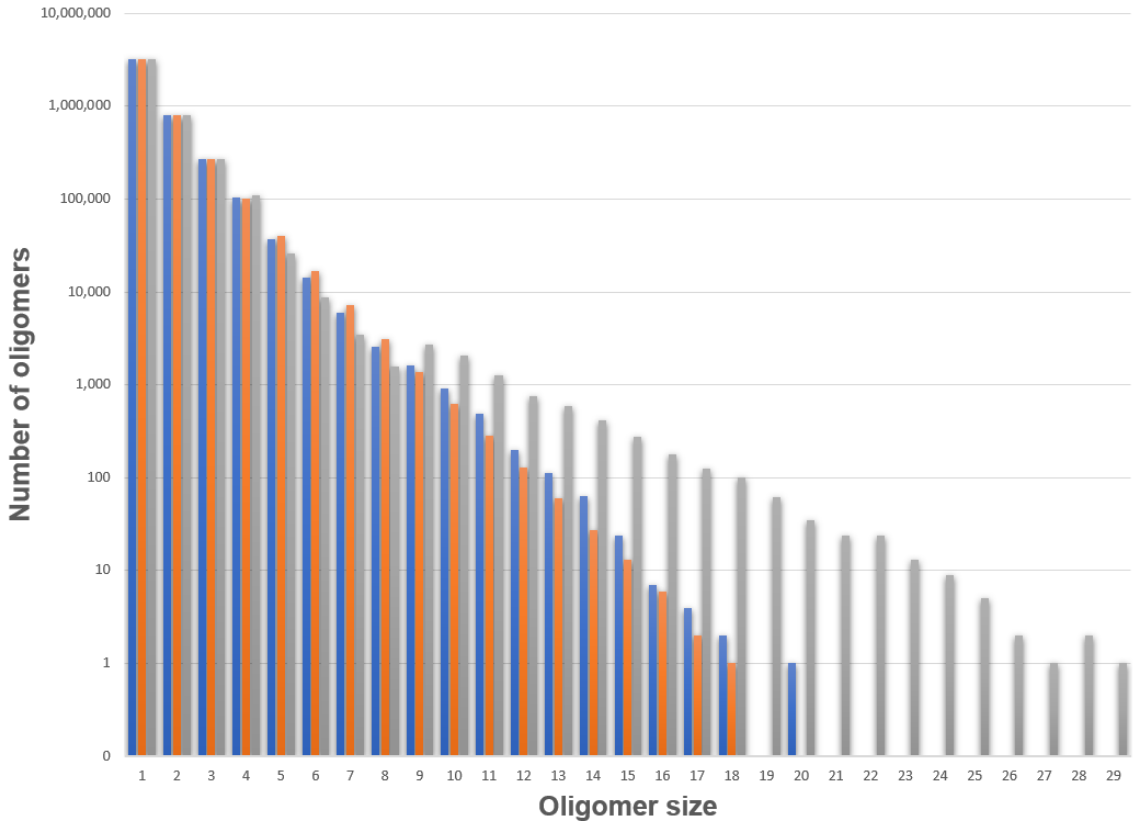

Figure 75: The exponential distribution of oligomers with structural assessment and ribozyme activity feeding back onto the recombination rate. The initial distribution is in orange, the unactivated distribution is blue, and the preactivated distribution is grey.

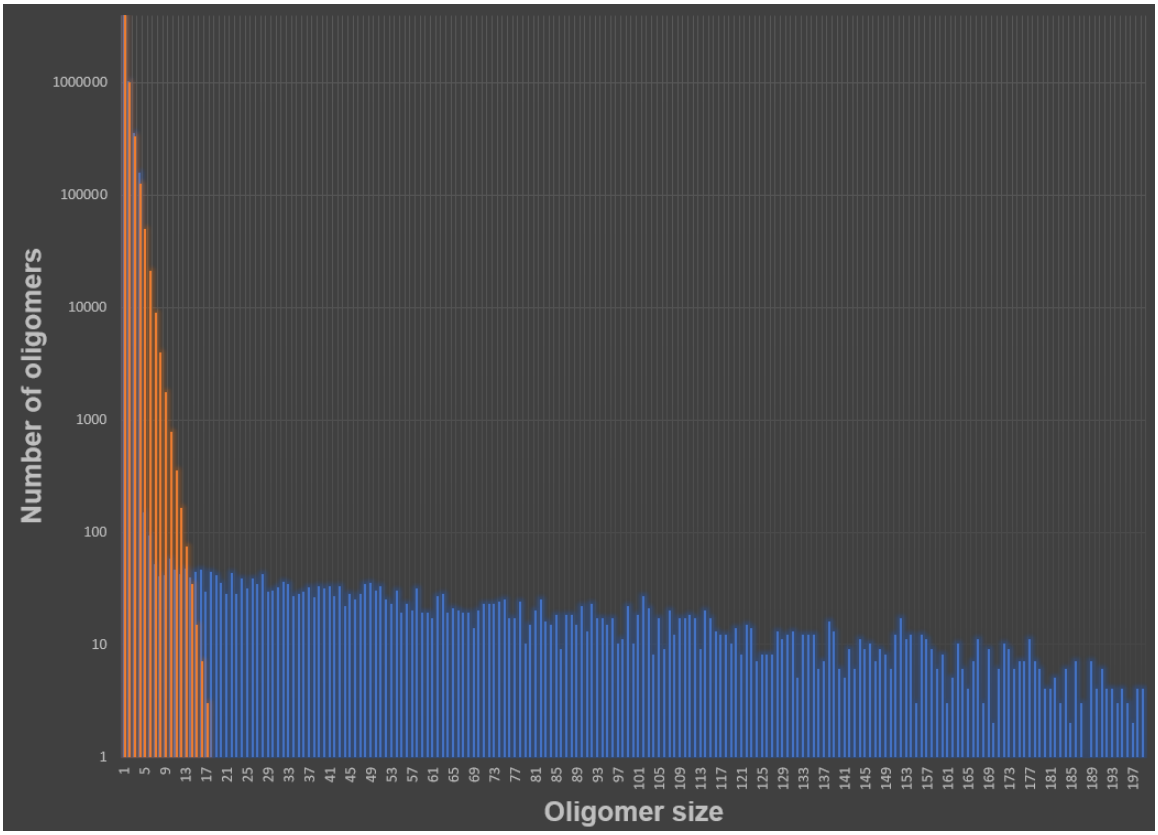

Figure 76: Redistribution of a pool of oligomers that is periodically replenished with fresh oligomers. All oligomers in the pool have a chance to be exchanged for fresh oligomers during replenishment events. The orange distribution is the initial distribution and the blue distribution is the final distribution.

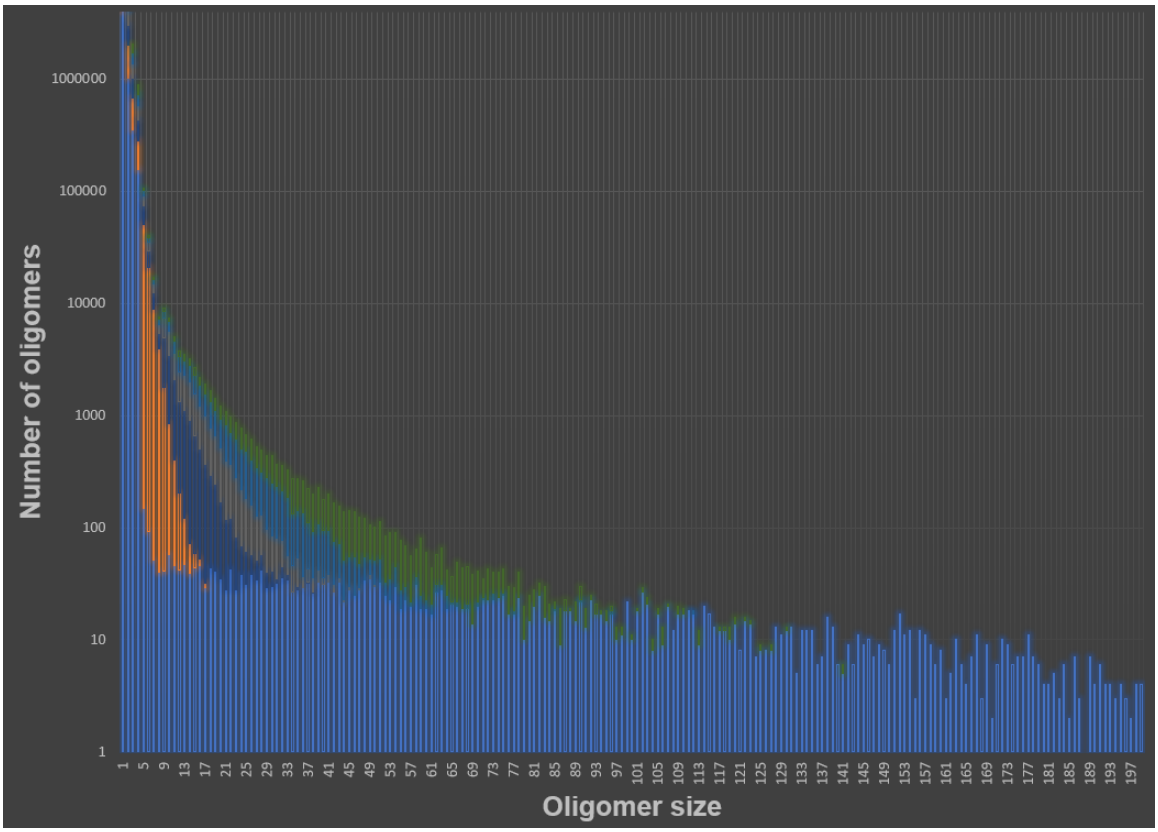

Figure 77: Timepoints of the replenished distribution. The peak distribution is at front in light blue. Orange represents the initial distribution. Dark blue is the distribution at 28 million generations, grey is the distribution at 28.4 million generations, light blue is the distribution at 28.8 million generations, and green is the distribution at 28.9 million generations.

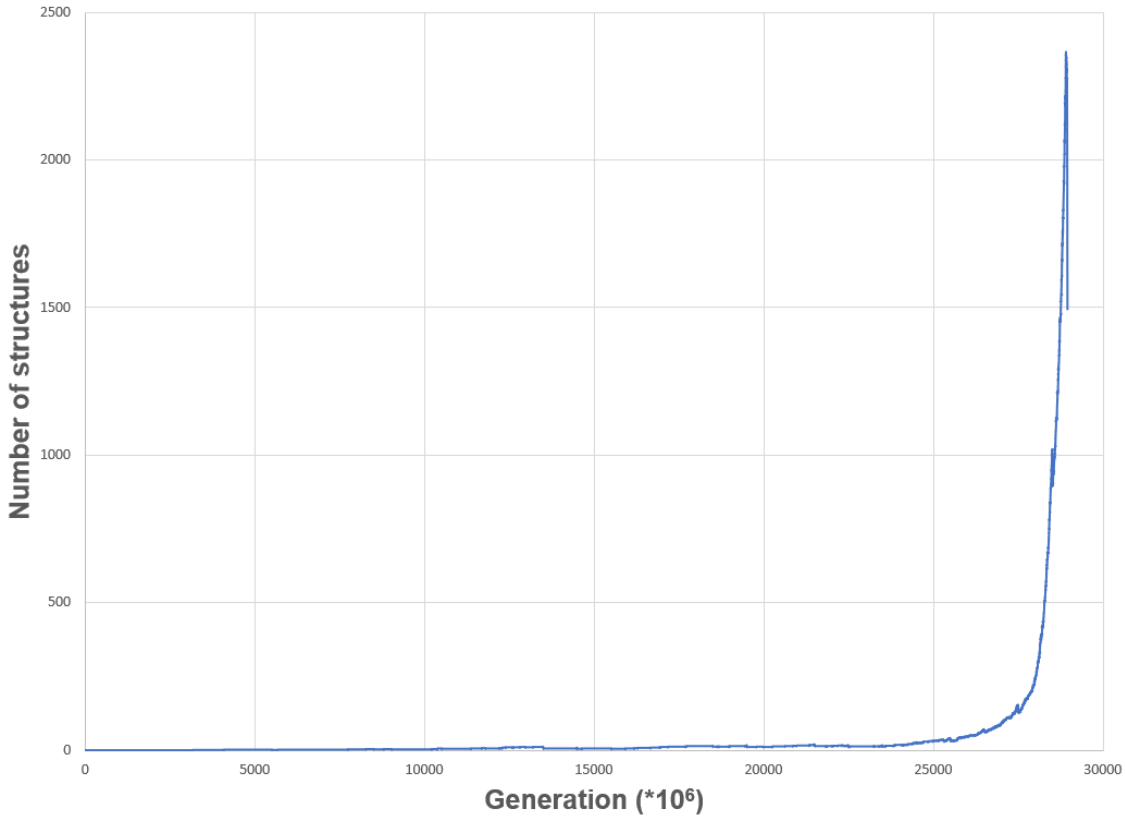

Figure 78: Structural content accumulating over time in the preactivated distribution with replenishment.

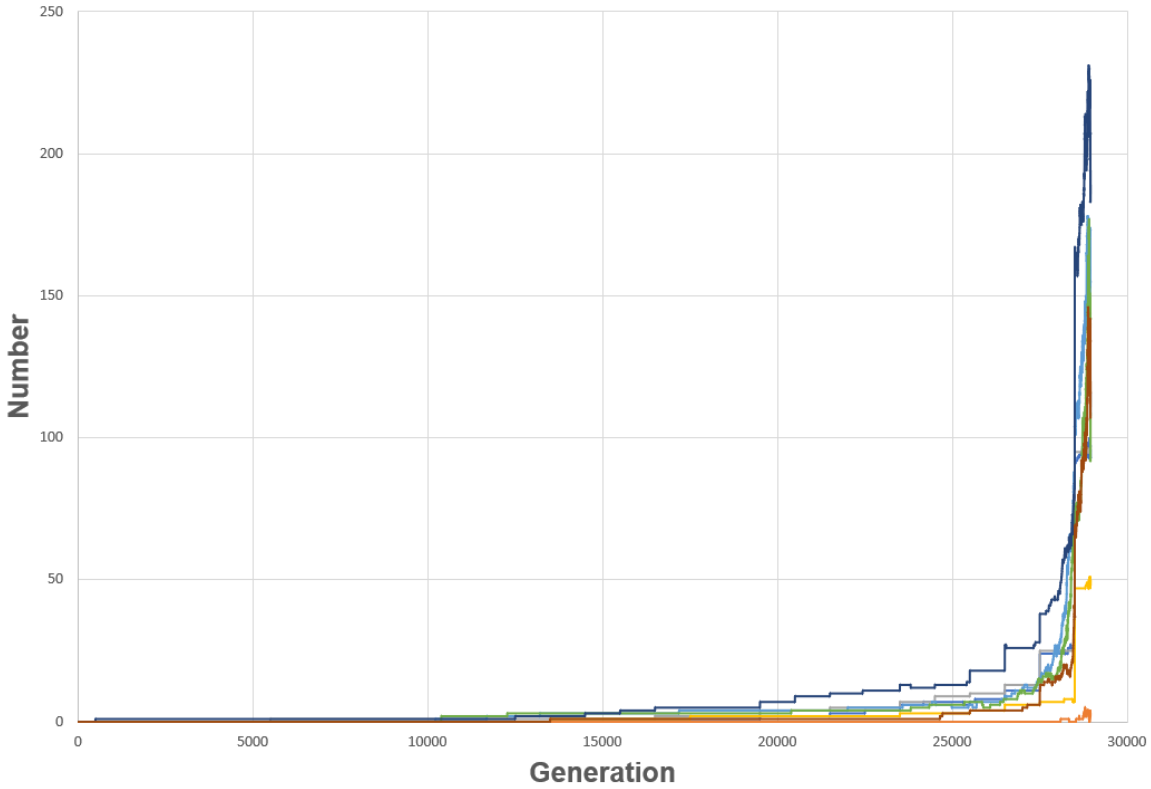

Figure 79: Types of structures generated in the replenished simulation over generational time. All of the different types of structures rapidly accumulate near the end. Blue represents GNRA tetraloops, orange UNCG, grey CUUG, yellow AANA, light blue generic tetraloops, green pentaloops, dark blue hexaloops, and brown represents heptaloops.

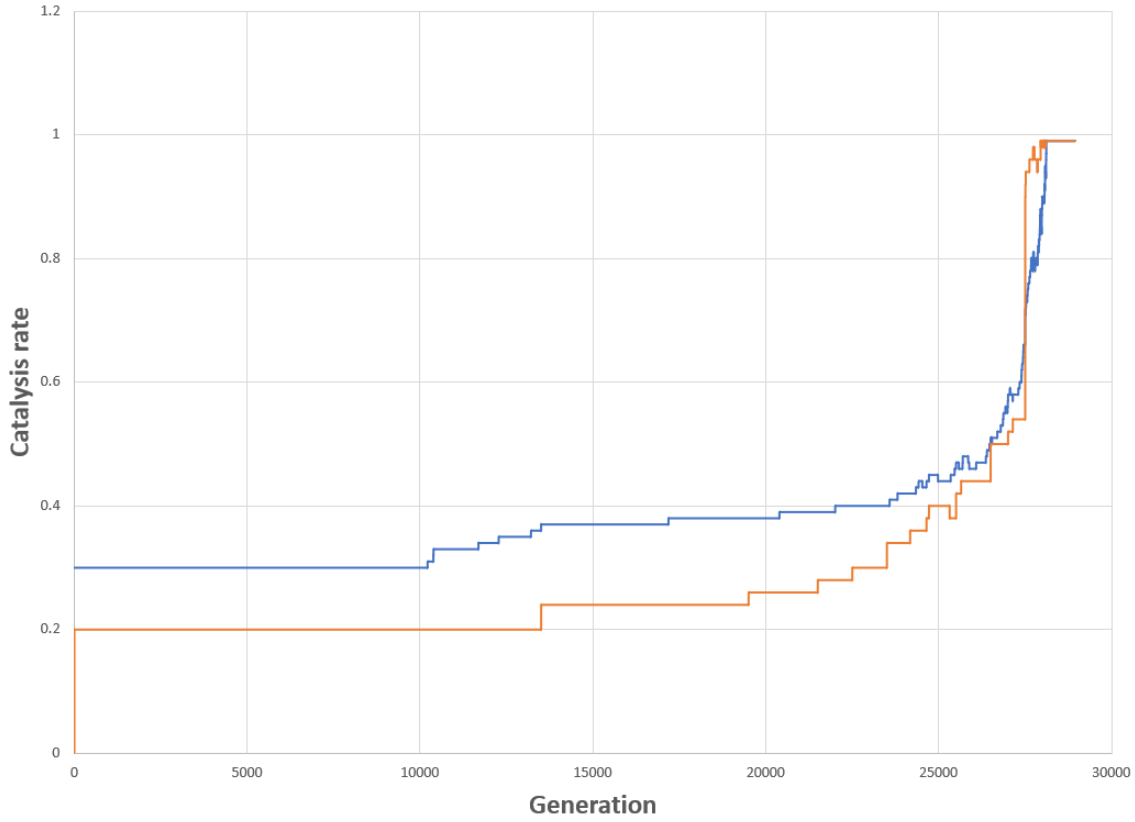

Figure 80: Catalysis rates over generations in the replenished simulation. The blue line represents the rate of cleavage, and the orange line is the rate of ligation. The x-axis “Generation” refers to millions of generations.

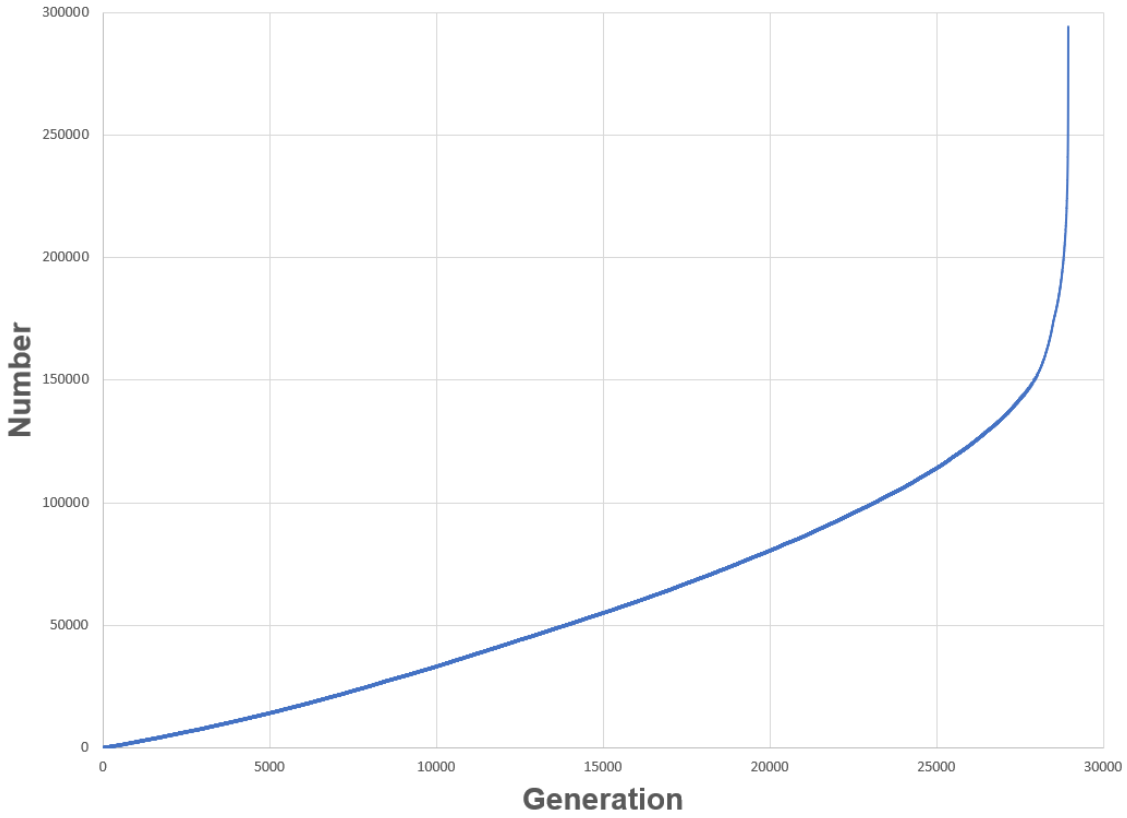

Figure 81: Cumulative alpha setups formed over time in the replenished simulation.

Chapter 5 Tables

Oligomer sequence	Initial amount	Final amount
GAGAGCAGGAA	3419	3419
CUCUCCUUCCUG	0	2239
CUCUCCUUCCUGCUCUCCUUCCUG	0	515
CUCUCCUUCCUGAAAA	6581	0
CUCUCCUUCCUGCUCUCCUUCCUGAAAA	0	966
CUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUG	0	81
CUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGAAAA	0	289
CUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUG	0	10
CUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGAAAA	0	46
CUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGAAAA	0	8
CUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGCUCUCCUUCCUGAAAA	0	1

Table 3: Simulated redistribution of oligomers from Lutay et al.[38], confirming that our simulated mechanism of alpha recombination is accurate. A majority of strands manifest as the cleavage product CUCUCCUUCCUG while the second most appear as the 28mer recombination product.

Unactivated				Activated		
Run	**Cleavages**	**Ligations**	**Recombinations**	**Cleavages**	**Ligations**	**Recombinations**
1	9049	338	2173	9573	5788	2368
2	8990	338	2231	9633	5748	2367
3	8923	356	2280	9536	5655	2322
4	8885	339	2245	9457	5580	2414
5	8888	362	2195	9603	5793	2489
6	9006	363	2221	9516	5886	2372
7	8914	392	2298	9576	5776	2389
8	8983	344	2258	9458	5626	2355
9	8940	325	2252	9552	5643	2437
10	8862	332	2258	9680	5677	2458
Avg	**8944**	**349**	**2241**	**9558**	**5717**	**2397**

Table 4: Average cleavages, ligations, and one-step recombinations over 10 simulations with either activated or unactivated pools.

Chapter 6: Conclusions

One of the biggest difficulties of many current origin-of-life scenarios is explaining how large catalysts based on polynucleotides or polypeptides can arise from essentially random sequences. It has never been clear if the origin of life was jumpstarted by the sudden emergence of significant catalysts, or if consequential catalysts evolved themselves by some kind of non-Darwinian chemical evolution. We have proposed that RNA recombination is a means by which RNAs can exchange genetic information in an energetically neutral fashion that does not require an advanced catalyst, and that this process is applicable even to small RNAs that could have been generated spontaneously by abiotic processes on the primordial earth. As such, the principal motivation and goal for our work in studying RNA recombination has been to find mechanisms by which RNAs can expand their lengths. To this end, we have demonstrated that mixtures of random RNAs are capable of undergoing recombination in slightly basic conditions and with magnesium to produce a distribution of larger products. The overall yield and reaction of such recombination reactions is small due in part to the reversibility of transesterification and the presence of competing degradation pathways, but the yield is significant enough to have implications for the origin of life. Our results have been further supported by the work of Mutschler et al.[51], who found nearly identical results with pools of random 20mers that were both activated and unactivated, although it is unclear what specific mechanisms are operative in random pools.

Here, we have used the individual oligomers **R16** and **H13**, which are 16 and 13 nucleotides long respectively, to demonstrate that there are verifiable mechanisms for the recombination of short RNAs. When incubated in the presence of magnesium at pH 8.0-

9.5 and cold cycling from 0°C to 22°C, both **R16** and **H13** undergo site-specific cleavage and ligation in a self-templating triplex to form products ranging from 28-30 nucleotides or 24-25 nucleotides respectively. The mechanism involves attack of the terminal 5′ hydroxyl of one strand onto the 2′-3′ cyclic phosphate of another strand that is left behind after the cleavage reaction. The overall mechanism is similar to the mechanism proposed and studied by Lutay et al.[38], but we have produced a result with oligomers containing 50% G-C and A-U content, and our reaction likely requires a bulge over which recombination takes place. We have termed this reaction the alpha-prime reaction, and to our knowledge it is the first ever demonstration of a self-templating and self-recombining RNA sequence at its size. We further generalized the reaction to all similar 16mers, and found two similar oligomers that undergo recombination. These studies also give further credence to the idea that recombination may be particularly facile over a three-nucleotide bulge.

A second mechanism we studied is the formation of a 31mer product that manifests solely in the **R16** self-reaction which we termed beta. Unlike the alpha-prime reaction, our HTS results implicate a molecule of **R16** whose 3′ end has been extended by 15 nucleotides. Here, we proposed that the beta reaction is the result of a two-step mechanism in which one internal 2′ hydroxyl of R16 attacks itself to form a lariat intermediate; re-attack of the 2′-5′ lariat branch by the terminal 3′ hydroxyl of another molecule of **R16** then produces the 31-nt product.

A third mechanism that we have proposed and provided evidence for is a branching reaction that can happen in both of the **R16** and **H13** self-reactions. In this case, the attack of an internal 2′-OH of one oligomer onto another may displace the 5′

piece of one RNA to form a 2′-5′ branch. We have been able to show that there is a fast and non-recursive reaction for both **R16** and **H13** that produces two general regions of products migrating more than twice the size of the starting material; we termed these regions gamma and gamma-prime. The slow migration of these products indicates nonlinear RNA; we furthermore do not observe any sequencing products above 31 nucleotides despite the fact that these bands are possibly the most intense and provide the highest yield of any other product in the self-reactions.

We examined the possibility of branched RNA with RNase digestions, and, critically, with 3′ adapter ligation to study migration differences. We found that 3′ adapter ligation generates an array of products whose total sum exceeds the number of products from the **R16** reaction. As no ligations happen in any of the controls, we conclude that these adapter ligations must solely result from attachment of the adapter to a free 3′ end of the product oligomers, and that the number of adapter products indicates that some of the recombinant products must have more than one 3′ end, providing good evidence for the presence of branched RNAs. However, despite the use of 2′ internal deoxy substitutions for all but three of the internal hydroxyls of **R16**, we were never able to conclusively eliminate any of the bands corresponding to the putative branched RNAs. In addition, an all-DNA version of R16 produces a single band in a similar region at more than twice the size of the starting material with no obvious intermediates, so it is possible that there is a fourth mechanism at play in which the RNA facilitates attack of a nitrogen atom from one of the bases onto a phosphodiester bond to form a branched product with a phosphoramidate linkage.

Given the sizes of the **R16** gamma products, we conclude that there must be a junction subject to attack by multiple internal hydroxyls and that attack of the same junction by different 2′ hydroxyls produces branches similar to others, accounting for the close spacing of the bands. Because of the close spacing of gamma products it is also difficult to detect if bands are lost when deoxyribose is substituted for ribose in various internal nucleotides. Future work to elucidate the specific structures of the products could include improved or altered reverse transcription to study the sequences of the branches, or mass spectrometry if enough product could be generated from a large-scale reaction.

Finally, we have constructed a sequence-explicit computer simulation of a pool of random RNAs undergoing iterative alpha recombination. In conjunction with our *in vitro* results, we have shown that alpha-recombination will cause a distributional shift from short oligomers to slightly longer ones, with the consequence of increased formation of even smaller oligomers and nucleotides that are the leftover pieces of disproportionation reactions. We have found that in all cases preactivation of a pool of RNA oligomers to facilitate ligation is a substantial accelerant for the upwards distributional shift of recombinant products, and that this effect is dampened by high rates of cleavage. We have further found that this distributional shift can occur in both flat distributions where the oligomers are all the same size, and exponential distributions, in which the number of RNAs at each size increases as the size gets smaller. Lastly, we conclude that if there is some form of replenishment, or some ability of recombined RNAs to persist in a pool while other less structured or smaller RNAs can be exchanged with the environment for new material, recombination can systematically expand the lengths of RNAs in a dynamic pool to build up a steady state distribution that includes RNAs of the sizes

necessary to be significant prebiotic catalysts or ribozymes, such as an RNA polymerase ribozyme.

Taken as a whole, our results have significant implications for the origin of life. The first and most significant finding is that RNAs of any size can undergo spontaneous recombination to form strands of longer sizes. Even if the production of long strands is directly offset by the generation of shorter strands, it is inevitable that a large amount of genetic scrambling may take place in a pool of homogenous, or possibly even heterogenous oligomers. This energetically neutral scrambling process appears to be a fundamental property of RNA, and could well be a phenomenon that contributes to the formation of critical RNA catalysts in any pool of oligomers. Scrambling is especially important in the context of sequence space, because as length increases, sequence space quickly becomes vast. It is not clear that any spontaneous polymerization of oligomers would in fact generate complete random sets of short oligomers such as 20mers, but even if it did, it would still be necessary to find mechanisms of exploring sequence space at larger sizes with less and less energetic expense[65]. In this way, recombination is far more viable as a prebiotic mechanism of sequence space exploration than abiotic polymerization as the length of oligomers increases – for example, even if mechanisms of abiotic polymerization were discovered that could generate randomers of sizes up to 100, it would not be possible to generate a complete set of such sizes due to the sheer mass of material required to form it.

A second critical feature that we have found with our results is that recombination, both *in vitro* and *in silico*, produces strands with considerably more structure than the starting material. In some respect, this is unsurprising, as longer

oligomers are more likely to have structure, but in our computer simulations it also appears that recombinant strands themselves have a greater chance of recombining than non-recombinant strands, likely because they possess intrinsic structure necessary to undergo recombination. This also has implications for sequence space exploration because any system of RNA undergoing length expansion by recombination is, on average, likely to have more structure than it would if the strands were randomly generated by abiotic processes, and structure is essential for catalysis.

Crucially, we have also found that recombination can take place in days or even hours in the presence of magnesium, slightly basic pH, and relatively benign conditions of modern room temperature, with slightly colder temperatures near the freezing or eutectic point appearing to be beneficial. This is important in terms of a time frame at the origin of life, since it is likely that life probably arose within 400 million years of Earth's formation and any process that could both spontaneously and efficiently explore sequence space would be essential to form life-like systems that relied on catalysis.

Finally, our work lays out a path for future experimentation on the recombination of short oligomers to expand their lengths in putative prebiotic conditions. The cleavage and ligation scheme we have designed, featuring recombination over a bulge, may or may not be specific to the nucleotides involved. It should be possible to optimize the scheme, identifying optimal nucleotides in the bulge, as well as finding the minimal length necessary to effect recombination. If this could be done, it may be possible to use the alpha-prime reaction to piece together larger RNA molecules in an energetically neutral fashion. Such work could further demonstrate the possibility of the emergence of a large catalyst solely by means of recombination.

One interesting question about our results is whether it is possible to increase the yield of the reaction. In terms of the origin of life, it is important that reaction yields be large enough to be consequential but small enough to maintain diversity. A pool of random or heterogenous RNA strands that strongly favored very specific reactions or types of reactions would result in a pool of more homogenous products. On the other hand, if every possible reaction in a random pool had at most a 1-2% yield, this would give sufficient diversity for the spontaneous emergence of catalysts with progressively better activities. Nonetheless, for quality *in vitro* results, it may be important to demonstrate increased yield. One idea in this regard is whether the replenishment addressed in our simulations can be carried out in a similar manner with real RNAs; if a stock of RNA can be periodically "refreshed" with new amounts of material, would the overall production of recombinant oligomers increase? If it could, it would be a demonstration that RNAs in some environmental gradient or compartment that retained molecules while allowing new ones in could have initiated the origin of life through recombination.

As a final roadmap for future experimentation on recombination, we propose the use of different pools of material and a more serious examination of structure present in recombinant products. Our work was conducted on both random pools of oligomers and homogenous pools of specific sequences. However, because random pools have an innate propensity to self-inhibit due to the presence of complementary strands, one possibility for future experimentation would be to use semi-random pools of various designs. The use of semi-random pools would substantially eliminate self-inhibition and provide a way of identifying the sequences of starting material in sequenced recombinant products,

allowing the possibility of finding new mechanisms or structures that facilitated small RNA recombination. Structural analysis of *in vitro* RNA recombination is also an endeavor that requires critical exploration. In our laboratory work, we have only given limited consideration to the structure of recombinant products, identifying their likely structures by sequencing, modeling, and considering whether the phosphor-ester bond formed during recombination is of the 2′-5′or 3′-5′ variety. Our simulation suggests that structure should be widespread in recombinant products. Therefore, it would be important to do a real structural analysis of the recombinant products, not only of homogenous pools like our work here, but of semi-random pools, and any mixtures where the predominant sequences can be identified or isolated, as well as assessment of branched, circular, and other nonlinear RNA structures that would indicate additional mechanisms of transesterification.

With our demonstration that small RNAs can undergo recombination to form longer products, and that incubating random RNAs that forms longer products over time, it has become possible to construct a narrative of the origin of life with recombination as the key to length expansion and complexity. That narrative runs something like this: In the beginning, there were small RNA oligomers formed by some abiotic or mineral catalyzed processes. These very short RNAs would not have exceeded 15–20 nucleotides in length and would not by themselves be significant catalysts. However, a general process of recombination over time could have increased the size and structural diversity of early RNA pools until there were fragments large enough to assemble into *trans* complexes or transient, catalytic tertiary structures[33]. These molecules could then have been the first enzymes. Some of these molecules would have had recombinase activity that could

accelerate the process of forming covalent bonds between *trans* complexes or even shorter RNAs. Recombinases that acted only at short, sequence-specific junctions could have recombined a wide variety of substrates, providing a mechanism of generating diversity. Constant turnover from recombination events and catalyzed cleavage and hydrolysis would have eliminated unstable structures and favored the formation of stable structures, leading to a gradual ecological succession of stable molecules and their component precursors. Finally, some of the larger products of this recombination explosion would have had real and powerful catalytic activity. This could have come in the form of an RNA polymerase ribozyme capable of synthesizing full-length copies of itself or its component precursors (which would recombine into the full-length polymerase ribozyme). Or, it could have been a primitive peptidyl-transferase center that polymerized short amino acid chains; these polypeptides could have become primitive RNA polymerases. Brought in proximity, a primitive peptidyl-transferase ribozyme and a polypeptide RNA polymerase could have formed an irreversible wheel that accelerated diversity, length, and activity until they could finally copy each other and form the ribonucleoprotein world, an obvious precursor to the last universal common ancestor.

References

1. Goldenfeld, Nigel. (2014) “Looking in the right direction: Carl Woese and evolutionary biology.” *RNA Biol.* 11(3): 248-253.
2. Weiss, M.C.; Preiner, M.; Xavier, J.C.; Zimorski, V.; Martin, W.F. (2018) “The last universal common ancestor between ancient Earth chemistry and the onset of genetics.” *PLoS Genet.* 14(8): e1007518.
3. Poole, A.M.; Horinouchi, N.; Catchpole, R.J.; Si, D.; Hibi, M.; Tanaka, K.; Ogawa, J. (2014) “The Case for an Early Biological Origin of DNA.” *J. Mol. Evol.* 79: 204.
4. Forterre, P.; Filée, J.; Myllykallio, H.; (2013) “Origin and Evolution of DNA and DNA Replication Machineries.” In: Madame Curie Bioscience Database [Internet]. Austin (TX): *Landes Bioscience*; 2000-2013.
5. Nissen, P., Hansen, J., Ban, N., Moore, P. B., Steitz, T. A. (2000) “The structural basis of ribosome activity in peptide bond synthesis.” *Science*, **289,** 920–930.
6. Brown, F.; Hull, R.; (1973) “Comparative Virology of the Small RNA Viruses.” *J. Gen. Virol.* 20: 43-60.
7. Burchell, M.J. (2004) “Panspermia today.” *Int. J. Astrobiol.* 3(2): 73-80.
8. Pressman, A.; Blanco, C.; Chen, I.A. (2015) “The RNA World as a Model System to Study the Origin of Life.” *Curr. Biol*. 25(19): PR953-R963
9. Orgel, Leslie. (2004) “Prebiotic Chemistry and the Origin of the RNA World.” *Crit. Rev. Biochem. Mol. Biol.* 39: 99-123.
10. Robertson, M.; Joyce, G.F. (2012) “The Origins of the RNA World.” *Cold Spring Harb. Perspect. Biol.* 4(5): a003608.
11. Ferré-D’amaré, A.R.; Zhou, K.; Doudna, J.A. “Crystal structure of a hepatitis delta virus ribozyme.” *Nature*. 1998, 395:567-574.
12. Ferré-D’amaré, A.R.; Scott, W.G.; (2010) “Small self-cleaving ribozymes.” *Cold Spring Harb. Perspect. Biol.* 2(10): a003574.
13. Gilbert, W. (1986) Origin of life: The RNA World. *Nature*, **319,** 618.
14. Attwater, J.; Holliger, P. (2014) “A synthetic approach to abiogenesis.” *Nat. Methods.* 11(5): 495
15. Kruger, K.; Grabowski, P.J.; Zaug, A.J.; Sands, J.; Gottschling, D.E.; Cech, T.R. (1982) “Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena.” *Cell*. 31(1): 147-157.
16. Zaug, A. J., Cech, T. R. (1986) The intervening sequence RNA of Tetrahymena is an enzyme. *Science*, **231,** 470–475.
17. Been, M.D.; Cech, T.R. (1988) “RNA as an RNA Polymerase: Net Elongation of an RNA Primer Catalyzed by the *Tetrahymena* Ribozyme.” *Science*, **239**, 1412-1416.
18. Bartel, D.P.; Doudna, J.A.; Usman, N.; Szostak, J.W. (1991) “Template-Directed Primer Extension Catalyzed by the *Tetrahymena* Ribozyme.” *Mol. Cell. Biol.* 11(6): 3390-3394.
19. Bartel, D.P.; Szostak, J.W. (1993) “Isolation of New Ribozymes from a Large Pool of Random Sequences.” *Science*, **261**, 1411-1418.

20. Horning, D.P.; Joyce, G.F. “Amplification of RNA by an RNA polymerase ribozyme.” (2016) *PNAS.* 113(35): 9786-9791.
21. Attwater, J.; Wochner, A.; Holliger, P. (2013) “In-ice evolution of RNA polymerase ribozyme activity.” *Nat. Chem*. 5(12): 1011-1018.
22. Ekland, E.H.; Bartel, D.P. (1995) “The secondary structure and sequence optimization of an RNA ligase ribozyme.” *Nucleic Acids Res*. 23(16): 3231-3238.
23. Ferris J.P.; Hill, A.R.; Jr, Liu R.; Orgel, L.E. (1996) “Synthesis of long prebiotic oligomers on mineral surfaces.” *Nature* **381**:59–61.
24. Gibard, C.; Bhowmik, S.; Karki, M.; Kim, E-K.; Krishnamurthy, R. (2018) “Phosphorylation, oligomerization and self-assembly in water under potential prebiotic conditions.” *Nat. Chem.* **10**:212–217.
25. De Guzman, V.; Shenasa, H.; Vercoutere, W.; Deamer, D. (2014) “Generation of oligonucleotides under hydrothermal conditions by non-enzymatic polymerization.” *J. Mol. Evol.* **78**:251–262.
26. Rajamani, S.; Vlassov, A.; Benner, S.; Coombs, A.; Olasagasti, F.; Deamer, D. (2008) “Lipid-assisted synthesis of RNA-like polymers from mononucleotides.” *Orig. Life. Evol. Biosph.* **38**:57–74.
27. Mutschler, H.; Holliger, P. (2014) “Non-canonical 3′-5′ extension of RNA with prebiotically plausible ribonucleoside 2′-3′ cyclic phosphates.” *J. Am. Chem. Soc.* 136(14): 5193-5196.
28. Prywes, N.; Blain, J.C.; Del Frate, F.; Szostak, J.W. (2016) “Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides.” *eLife* **5:**e17756.
29. Müller, S.; Appel, B.; Krellenberg, T.; Petkovic, S. (2012) “The Many Faces of the Hairpin Ribozyme: Structural and Functional Variants of a Small Catalytic RNA.” *Life*. 64(1): 36-47.
30. Hamann, C.; Luptak, A.; Perreault, J.; de la Peña, M.; (2012) “The ubiquitous hammerhead ribozyme.” *RNA*. 18(5): 871–885.
31. Popovic, M.; Ditzler, M. “Molecular crowding and evolution of ligase ribozymes.” (2017) *Astrobiology Science Conference,* contribution 1965. (https://www.hou.usra.edu/meetings/abscicon2017/pdf/3734.pdf)
32. Turk, R.M.; Chumachenko, N.V.; Yarus, M. (2010) “Multiple translational products from a five-nucleotide ribozyme.” *PNAS.* 107(10): 4585-4589.
33. Doudna, J.A.; Cech, T.R. (1995) “Self-assembly of a group I intron active site from its component tertiary structural domains.” *RNA,* **1,** 36–45.
34. Hayden, E.J.; Lehman, N. (2006) “Self-assembly of a group I intron from inactive oligonucleotide fragments.” *Chem. Biol*. **13,** 909–918.
35. Cech, T.R. (1990) “Self-splicing of Group I Introns.” *Annu. Rev. Biochem.* 59: 543-568.
36. Scott, W.G.; Szöke, A.; Blaustein, J.; O’Rourke, S.M.; Robertson, M.P. (2014) “RNA Catalysis, Thermodynamics, and the Origin of Life.” *Life(Basel).* 4(2): 131–141.
37. Pino, S.; Costanzo, G.; Giorgi, A; Šponer, J.; Šponer, J.E.; Di Mauro, E. (2013) “Ribozyme activity of RNA nonenzymatically polymerized from 3′,5′-cyclic

GMP." *Entropy* **15:**5362–5383.
38. Lutay, A.V.; Zenkova, M.A.; Vlassov, V.V. (2007) "Nonenzymatic recombination of RNA: possible mechanism for the formation of novel sequences." *Chem. Biodivers*. **4,** 762–767.
39. Smail, B.A.; Clifton, B.; Mizuuchi, R.; Lehman, N. (2019) "Spontaneous advent of genetic diversity in RNA populations through multiple recombination mechanisms." *RNA*. 25:453-464.
40. Chen, Y.; Zheng, Li.; Liu, B.; Zhong, S.; Giovannoni, J.; Fei, Z. (2012) "A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA Ligase 1 adenylated adapters." *Plant Methods*. 8:41.
41. Coppins, R.L.; Silverman, S.K. (2004) "A DNA enzyme that mimics the first step of RNA splicing." *Nat. Struct. Mol. Biol*. **11**(3):270-274.
42. Pratico, E.D.; Wang, Y.; Silverman, S.K. (2005) "A deoxyribozyme that synthesizes 2′-5′ branched RNA with any branch-site nucleotide." *Nucleic Acids Res.* **33**(11):3503-3512.
43. Imburgio, D.; Rong, M.; Ma, K.; McAllister, W.T. (2000) "Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants." *Biochemistry*. **39**:10419-10430.
44. Kao, C.; Zheng, M.; Rüdisser, S. (1999) "A simple and efficient method to reduce nontemplated nucleotide addition at the 3′ terminus of RNAs transcribed by T7 RNA polymerase." *RNA*. **5**(9):1268-1272.
45. Zuker, M. (2003) "Mfold web server for nucleic acid folding and hybridization prediction." *Nucleic Acids Res.* 31(13): 3406-3415.
46. Janssen, Stefan.; Giegerich, Robert. (2015) "The RNA shapes studio." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu649
47. Usher, D.A.; McHale, A.H. (1976) "Hydrolytic stability of helical RNA: A selective advantage for the natural 3′-5′ bond." *PNAS*. 73(4): 1149-1153.
48. Liu, J.; Lilley, D. M. J. (2007) "The role of specific 2′ hydroxyl groups in the stabilization of the folded conformation of kink-turn RNA." *RNA*. 13(2): 200-210.
49. Poudyal, R.; Phuong, D. M. N.; Lokugamage, M. P.; Callaway, M. K.; Gavette, J. V.; Krishnamurthy, R.; Burke, D. H. (2017) "Nucleobase modification by an RNA enzyme." *Nucleic Acids Res.* 45(3): 1345–1354.
50. Suzuki, H.; Zuo, Y.; Wang, J.; Zhang, M.; Malhotra, A.; Mayeda, A. (2006) "Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing." *Nucleic Acids Res*. 34(8): e63.
51. Mutschler, H.; Taylor, A. I.; Porebski, B.; Lightowlers, A.; Houlihan, G.; Abramov, M.; Herdewijn, P.; Holliger, P. (2018) "Random-sequence genetic oligomer pools display an innate potential for ligation and recombination." *eLife*. 7:e43022
52. Boerlijst, M.C.; Hogeweg, P. "Self-structuring and selection: Spiral waves as a substrate for prebiotic evolution, ed. by C.G. Langton, C. Taylor, J.D. Farmer, S. Rasmussen, *Artificial Life II* (Addison Wesley, 1991), pp. 255–276
53. Dyson, F.J. "Origins of Life." 2nd edition. 1999, Cambridge University Press, New York.

54. Szabo, P.; Scheuring, I.; Czaran, T.; Szathmary, E. (2002) “In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity.” *Nature*. 420:340-343.
55. Wu M.; Higgs P.G. (2009) “Origin of Self-replicating Biopolymers: Autocatalytic Feedback can Jump-start the RNA World.” *J. Mol. Evol.* 69: 541-554.
56. Wu, M.; Walker, S.I.; Higgs, P.G. (2012) “Autocatalytic Replication and Homochirality in Biopolymers: Is Homochirality a Requirement of Life or a Result of It?” *Astrobiology*. 12 (9) 818-829.
57. Higgs, P.G. (2016) “The Effect of Limited Diffusion and Wet-Dry Cycling on Reversible Polymerization Reactions: Implications for Prebiotic Synthesis of Nucleic Acids.” *Life.* 6(2):24.
58. Wu, M.; Higgs, P.G. (2008) “Compositional Inheritance: Comparison of Self-assembly and Catalysis.” *Orig. Life. Evol. Biosph*. 38:399-418.
59. Wu, M.; Higgs, P.G. (2011) “Comparison of the Roles of Nucleotide Synthesis, Polymerization, and Recombination in the Origin of Autocatalytic Sets of RNAs.” *Astrobiology*. 11(9): 895-906.
60. Ma, W.; Yu, C.; Zhang, W. (2007) “Monte Carlo simulation of early molecular evolution in the RNA World.” *Biosystems.* 90:28-39.
61. Ma, W.; Yu, C.; Zhang, W.; Zhou, P.; Hu, J. (2011) “Self-replication: Spelling It out in a Chemical Background.” *Theory. Biosci.* 130(2):119-25.
62. Ma, W.; Yu, C.; Zhang, W.; Hu, J. (2007) “Nucleotide synthetase ribozymes may have emerged first in the RNA World.” *RNA.* 13:2012-2019.
63. Mariani, A.; Russell, D.; Javelle, T.; Sutherland, J. D. (2018) “A Light-Releasable, Potentially Prebiotic Nucleotide Activating Agent.” *J. Am. Chem. Soc.* 140:28.
64. Myszor, D.; Cyran, K.A. (2010) “Influence of non-enzymatic template-directed RNA recombination processes on polynucleotides lengths in Monte Carlo simulation model of the RNA World.” *JAMI.* 4:676-681.
65. Blokhuis, A., Lacoste, D. (2017) Length and sequence relaxation of copolymers under recombination reactions. *J. Chem. Phys*. 147:094905